ICES Journal of Marine Science
November 2011, Volume 68, Issue 10, Pages 2086-2095
http://dx.doi.org/10.1093/icesjms/fsr139

Spatially explicit estimation of fish length histograms, with application to anchovy habitats in the Bay of Biscay

Pierre Petitgas*, Mathieu Doray, Jacques Massé and Patrick Grellier

Ifremer, Department of Ecology and Models for Fisheries, rue de l'Île d'Yeu, BP 22105, 44300 Nantes, France

*: Corresponding author : Pierre Petitgas, tel: +33 240 374163 ; fax: +33 240 374075 ;
e-mail address : pierre.petitgas@ifremer.fr

Abstract :

Fish length frequency histograms from research surveys are of prime importance for identifying habitats of different life stages, as well as for stock assessment. However, no method has thus far been available for mapping these histograms as spatially varying curves. Here, a procedure is applied to map spatially connected curves, and detail is given on how it can be applied to map the length frequency histograms. At each sample location, a fish length frequency histogram is given as a vector of non-independent values. The histogram is first modelled as a polynomial expansion on the basis of orthogonal polynomials. Then, the polynomial coefficients are mapped by co-kriging, after fitting a model of co-regionalization. The length frequency map is finally derived by linearly combining maps of polynomial coefficients. An estimation variance associated with the map is also derived. Maps of anchovy length distributions are produced by applying the method to midwater trawl length data from the PELGAS acoustic surveys in the Bay of Biscay. This novel approach extends the application of kriging techniques to curves or functions, opening new perspectives for mapping more complex information than just the values of fish density.

Keywords : co-kriging ; European anchovy ; fish length ; functional kriging ; habitat ; mapping curves

1. Introduction

Body size is an important ecological attribute of a fish, because length influences its ability to use habitats efficiently (Werner and Gilliam, 1983). Spatial segregation by length is common in marine fish populations with complex life cycles, where individual fish make ontogenic shifts in habitat (e.g. ICES, 2010). In this context, maps of fish length are useful for various purposes in fisheries science, including estimating population abundance by research survey, ecological understanding of life cycle patterns, spatial modelling of populations, and designing marine protected areas (MPAs). During demersal trawl surveys and pelagic acoustic surveys, abundances per species and length class in trawl samples are used routinely to derive estimates of fish population abundance at length and age. Fish length also plays a key role in the survey-based assessment of pelagic fish stocks by acoustic methods, because the target strength of an individual fish, i.e. its acoustic reverberation index, is a direct function of fish length. An acoustic estimate of fish abundance is in that case obtained by combining maps of mean fish length and acoustic backscatter (Woillez *et al*., 2009). Another important application of maps of fish length is for parametrizing spatially explicit population models. In such models, maps of fish length or age and habitat connectivity are in general parametrized on the basis of mean spatial patterns emerging by averaging length and age maps over many years (Christensen *et al*., 2009). MPAs are designed on the basis of such maps that characterize the habitats of particular life stages (Botsford *et al*., 2009). However, variation across the years in spatial distribution may jeopardize the effectiveness of an MPA (van Keeken *et al*., 2007), so a procedure is needed to map length distributions of fish and the associated estimation variances.

Fish length is often mapped as mean length (Chapter 4 of Rivoirard *et al*., 2001). In that case it is assumed that a fish length-frequency histogram is adequately represented by the mean value, i.e. that it is more or less unimodal. This assumption can be reasonable, because small and large fish usually occupy different habitats, but when several cohorts are found in the same area, such histograms will be either bimodal or more complex in shape. Another approach is to map fish density for particular length groups (van Keeken *et al*., 2007), targeting the mapping to a particular portion of the histogram. However, different length groups cannot be mapped coherently if they are mapped separately, because their coherent mapping requires the modelling of spatial cross-correlation between length groups. Co-kriging (Chapter 5 of Chilès and Delfiner, 1999) can therefore be both useful and necessary, because if allows coherent mapping across length groups. The methodology used is designed to map the length-frequency distribution as a continuous curve. It is flexible and applicable to many and varied types of distribution pattern. Although more complex than simple block-averaging the length distribution of the data (Supplementary material Figure S1**)**, the method has the advantages of kriging, which are (i) optimal interpolation at each point of a grid, and (ii) calculation of the corresponding estimation variance.

Contrary to the mapping of fish density or mean length, which consists of estimating one value at each grid node, mapping a fish length-frequency histogram requires the mapping of a function or curve, i.e. a vector of non-independent values. In functional data analysis (Ramsay and Silverman, 2005), functions are generally modelled and smoothed as linear combinations of basis functions. Nerini *et al*. (2010) developed a method for mapping spatially connected curves, which in practice amounts to (i) modelling the curves by polynomial expansion, and (ii) co-kriging the polynomial

coefficients. Here, we explore and discuss the applicability of this method to mapping fish length-frequency histograms from fisheries survey data. We then apply the methodology to a series of anchovy length measurements obtained from midwater trawls performed during acoustic surveys in the Bay of Biscay, to illustrate its potential for size-dependent habitat mapping. To our knowledge, this is the first time that this method has been applied to fisheries survey data and for such histograms to be mapped as continuous curves.

2. Material and methods

2.1. Data

Fish-length data were obtained by sampling the catches of midwater pelagic trawls during the PELGAS cruise, a pelagic acoustic survey conducted by Ifremer in the Bay of Biscay each spring since 2000. During the acoustic surveys, trawls were made on echotraces for identification purposes, and to collect fish for recording biological parameters (Petitgas *et al*., 2003a). Most fish echotraces were organized vertically, located between 0 and 50 m above the seabed. They were fished using a 25-m vertical opening midwater trawl towed close to the seabed. The mean haul duration was 45 min at 4 knots, yielding a trawled distance of 3 nautical miles (miles hereafter). The catch was sorted and weighed by species after each haul. A random subsample of each species was taken to establish length frequencies for each haul. The data were organized in matrix form, with stations as rows and length classes as columns. For each haul, the sampled length-frequency histogram was a vector of the proportions at length, which summed to unity. The position of a trawl was taken as the track midpoint. The target species was anchovy, and total length was measured to the nearest 0.5 cm below.

First, the entire set of 497 trawls performed over the period 2000–2010 was used to map the probability of anchovy presence. For that, at each haul location x, an indicator of anchovy presence $1(x)$ was defined as $1(x)=1$ if the anchovy catch was ≥ 2 kg, and $1(x)=0$ otherwise. Then, trawls with sufficient anchovy measurements to construct an anchovy length-frequency histogram were selected. These “positive” anchovy hauls needed at least 30 anchovy to be measured and a total anchovy catch of at least 2 kg. Using this criterion, we retained a total of 226 positive anchovy hauls from the surveys conducted between 2000 and 2010 (Figure 1).

The probability of anchovy presence over the series was mapped using the entire set of trawls. Then, the length frequency distribution using positive hauls only was mapped. Finally, we combined the map of the probability of anchovy presence with that of anchovy length frequency to access the spatial patterns of the “most probable” size-dependent potential habitats, i.e. habitats occupied by anchovy with sufficient probability during the 10 years survey series. In our analyses, time was collapsed by pooling the hauls from different surveys.

2.2. Mapping fish presence: indicator kriging

The indicator of anchovy presence was mapped by ordinary kriging in a moving neighbourhood (Chapter 3 of Chilès and Delfiner, 1999). The same grid was used as for fish length-frequency mapping. Ordinary kriging is a linear estimator in which the weights assigned to the samples are such that the estimator is unbiased and of minimal

variance. The procedure consists of (i) computing and modelling the variogram, $\gamma(h)$, and (ii) solving the ordinary kriging system at each grid node (see Appendix).

2.3. Mapping length frequencies: polynomial expansion and co-kriging

Let $y(i,t)$ represent a curve at location x_i, where t varies in the interval τ. Sampling provides information on the spatially connected curves y as a vector of n values at each sample location x_i, $i = \{1, \ldots, N\}$. The curves $y(i,t)$ can be expanded on a basis of orthogonal polynomials ϕ_k:

$$y(i,t) = \sum_{k=0}^{+\infty} z_k(i)\ \phi_k(t) \quad , \tag{1}$$

where $z_k(i)$ is the coefficient at location x_i of the polynomial of degree k and $\phi_k(t)$ is the value of the polynomial of degree k for parameter t (e.g. length). Nerini *et al*. (2010) showed that a geostatistical estimate of the curve $y(t)$ at unknown location x_0 can be obtained by co-kriging the z coefficients:

$$y^{\mathrm{CK}}(o,t) = \sum_{k=0}^{P} z_k{}^{\mathrm{CK}}(o)\ \phi_k(t)\,, \tag{2}$$

where CK refers to co-kriging, and function y is approximated by the polynomial expansion of order P. Polynomial expansion allows one to model the experimental vectors of observed values as smooth continuous curves. Co-kriging allows one to map the polynomial coefficients jointly, making use of their spatial cross-correlations. The procedure therefore yields a map of continuous curves.

Equation (2) was applied to estimate the proportions of fish at length t. Building on the fact that the co-kriging estimate of a linear combination of variables is the combination of their co-kriged estimates (Chapter 5 of Chilès and Delfiner, 1999), the map of the proportions of fish within a given length range $[t_1,t_2]$ can be derived from the maps of the co-kriged polynomial coefficients:

$$\left(\int_{t1}^{t2} y(x,t)dt\right)^{\mathrm{CK}} = \sum_{k=0}^{P} z_k{}^{\mathrm{CK}}(x)\int_{t1}^{t2}\phi_k(t)\mathrm{d}t\,. \tag{3}$$

Fitting an appropriate model of the spatial variation of curves $y(i,t)$ depends on the appropriate definition of (i) the order P of the polynomial expansion, and (ii) the spatial cross-correlations between polynomial coefficients. The order P depends on the curve complexity to be modelled: the higher the polynomial degree, the more complex the oscillations of y modelled. The spatial cross-correlations between coefficients will characterize how portions of the curve (for different values of t) co-vary at distance vector h apart, i.e. how small, medium, and large values of fish length are spatially cross-correlated. The spatial covariation of polynomial coefficients is used in the co-kriging procedure, resulting in a map of the fish length histograms and preserving the relative proportions between length classes (because they all sum to unity).

The steps of the method can be summarized as follows:

(i) select the orthogonal polynomial basis and define the order P of polynomial expansion to model the data curves at the sample locations;
(ii) at each sample location x_i, estimate the polynomial coefficients $z_k(i)$ $\{k = 0,\ldots,P\}$;
(iii) compute the (P+1, P+1) matrix of cross-variograms between the polynomial coefficients;
(iv) fit a linear model of co-regionalization to the multivariate dataset of polynomial coefficients (P+1 variables valued at N locations);
(v) carry out co-kriging of the polynomial coefficients;
(vi) estimate the histogram at each grid node using the polynomial expansion [Equation (2)] and derive the estimation variance.

It was deemed convenient to use normalized Legendre polynomials (Chapter 1 of Gautschi, 2004), which are orthonormal on the interval [–1,+1]: $\int_{-1}^{+1} \phi_k(t)\phi_q(t)\mathrm{d}t = 0$ for $k \neq q$.

The Legendre polynomials were computed by the three-term recurrence relation

$$\phi_0(t)=1;\ \phi_1(t)=t;\ (k+1)\phi_{k+1}(t)=(2k+1)t\phi_k(t)-k\phi_{k-1}(t)\ ;\ t\in[-1,+1];\ k\geq 2\ ,$$

and normalized by dividing by their norm, $\|\phi_k\| = \sqrt{2/(2k+1)}$.

The length values l in the interval $[l_1,l_2]$ were transformed into variable t in the interval $[-1,+1]$ as: $t=2(l-l_0)/(l_2-l_1)$; $l_0=(l_1+l_2)/2$. In the present study, the same length classes were measured at each station, so each curve $y(i,t)$ was valued at the same values of t, which corresponded to a sampling discretization of interval τ, irrespective of the sample location x_i. We therefore had N stations and the same n length classes at each station.

In steps (i) and (ii), polynomial expansions were fitted to histogram curves by least squares, with increasing polynomial degree. The polynomial degree retained was defined based on a goodness-of-fit criterion (gof) defined as

$$\mathrm{gof}(i,nk) = \sum_t \left[y(i,t) - \sum_{k=0}^{nk} \hat{z}_k(i)\ \phi_k(t) \right]^2 / \sum_t \left[y(i,t)\right]^2 , \qquad (4)$$

where i is the spatial location index and nk the degree of the polynomial expansion.

The gof includes a scaling factor in its denominator, allowing for comparison of the fits of different polynomial expansions to different curves y. As the degree nk increases, the gof decreases. The retained $nk = P$ degree was defined as the lowest degree for which the average gof over all stations i was <0.05. That value lay near the inflection point of the decreasing gof vs. nk curve, beyond which the decrease in gof slowed notably. The selected polynomial order induced a 5% reduction in the total variability of the experimental histograms when fitting the polynomial expansion.

In steps (iii)–(v), the polynomial coefficients $z_k(i)$, $i = \{1,\ldots N\}$, $k = \{0,\ldots,P\}$ formed a set of spatially covarying variables. Co-kriging was used to estimate each variable by taking into account its spatial correlation with all other variables (see Appendix). Polynomial coefficients were considered as quasi-stationary in space: their true (unknown) means were assumed to be constant or varying sufficiently smoothly to be considered as constant in restricted neighbourhoods around each grid point (ordinary kriging). In step (vi), the estimation variance of the polynomial expanded curve $y(0,t)$ [Equation (2)] was approximated by combining the polynomial co-kriging estimation variances (Appendix).

A linear model of co-regionalization (Chapter 23 of Wackernagel, 1995; Chapter 5 of Chilès and Delfiner, 1999) was used to co-krige the polynomial coefficients. This model required joint modelling of the direct and cross-variograms $\gamma_{kk'}(h)$ between all variables k and k', $k = \{0,\ldots,P\}$, $k' = \{0,\ldots,P\}$. In this versatile model, direct and cross-variograms are linear combinations of a small number s of elementary unit variance structures $\gamma_u(h)$: $\gamma_{kk'}(h) = \sum_{u=1}^{s} b_u(k,k')\,\gamma_u(h)$. Joint modelling of all (cross-)variograms was achieved using the algorithm of Goulard and Voltz (1992), which ensures that the model-based variances are non-negative through the inequality $|b_u(k,k')| \le \sqrt{b_u(k,k)\,b_u(k',k')}$, $u = \{1,\ldots s\}$. This inequality states that the elementary structures comprising the cross-variograms are those already in the direct variograms. Therefore, in practice, the elementary structures u are first identified in the direct variograms and subsequently used to model the cross-variograms. The algorithm of Goulard and Voltz (1992) then allows for joint fitting of all (cross-)variogram sills $b_u(k,k')$, $k = \{0,\ldots,P\}$, $k' = \{0,\ldots,P\}$, $u = \{1,\ldots s\}$.

Fitting the linear model of co-regionalization, co-kriging the polynomial coefficients and kriging the indicator of fish presence were performed using the R library RGeoS (Renard and Bez, 2008).

3. Results

3.1. Probability map of fish presence

The variogram of the indicator of fish presence (not shown) was isotropic and modelled as the sum of a nugget effect (sill = 0.145) and a spherical variogram (sill = 0.105, range 120 miles). The probability map of anchovy presence (Figure 2) showed a strong spatial pattern. Two habitats had a very high probability of fish presence: the shelf break close to 44.75°N 02°W, and an area off the Gironde estuary close to 45.5°N 01.5°W. These areas were contained inside a larger area south of 46.5°N and east of 2.5°W, where anchovy had a high probability of presence. Outside that rectangle, anchovy presence was sporadic. As all the data collected over the 10 years of research cruises were treated together, a low probability of anchovy presence might be attributable to either a sporadic occurrence across years or spatial heterogeneity, or both.

3.2. Polynomial expansion

Polynomial expansions were fitted to the sampled fish length-frequency histograms with increasing polynomial degree. The variability in the gof between stations for a fixed degree characterized the ability of the polynomial expansion to model the different

situations. We attempted to both minimize the average gof and the between-station gof variability. The gof decreased with increasing polynomial degree, rapidly until 10–13°, and more slowly beyond that (Figure 3). Interstation gof variability also decreased with increasing polynomial degree (Figure 3). However, some interstation variability was present at higher polynomial orders, meaning that the sample histograms were noisy at some stations. A too-smooth fit would jeopardize the mapping of rapid spatial transitions in the modes of the histogram. The polynomial degree retained was $P = 13$, corresponding to a gof of 0.05, which was a satisfactory compromise between a low gof and a smooth-enough fit of the histogram at all stations.

3.3. Co-regionalization model

The 14 polynomial coefficients z_k, $k = \{0, \ldots, 13\}$, formed a co-regionalized set of variables. Direct and cross-variograms were computed with a lag of 5 miles. No particular anisotropy was detected on direct variograms, so only omnidirectional cross-variograms were computed and modelled. Most of the direct variograms showed a nugget effect and a spherical component, and some a linear component with no obvious sill (Figure 4). All variograms were modelled with two structures ($u = \{1,2\}$): a nugget effect and a spherical model with a correlation range of 20 miles. The reason was that both linear and spherical models behave linearly for short distances, which applies within the kriging neighbourhoods used (see below). When the structure u (e.g. the correlation range) in the cross-variogram $\gamma_{ij}(h)$ between variables z_i and z_j was the same as on the direct variograms, the ratio $b_u(i,j)/\sqrt{b_u(i,i)b_u(j,j)}$ was close to –1 or +1 (negative or positive cross-correlation). Alternatively, when the structure on the direct variograms was not present on the cross-variogram, the ratio was close to 0 (no cross-correlation). The result of the automatic fit for each of the variance–covariance matrices b_u ($u = \{1,2\}$) can be seen by plotting the values in the plane (b_1,b_2) (Figure 5): cross-variograms displayed either both structures (points with abscissa and ordinate close to +1 or –1), only one structure (points along the x- or the y-axis), or no structure (points close to the origin). A few pairs of polynomials were highly correlated [i.e. close to (+1,+1), pairs 0–4, 2–6, 1–5], and a larger number were anti-correlated [close to (–1,–1)], and most pairs were scattered around the origin, meaning that they displayed the two structures, but with small or medium coefficients. Examples of different cross-variogram fits are given in Figure 6.

3.4. Co-kriging and mapping the histogram

The polynomial coefficients were estimated by ordinary co-kriging at each point of a regular grid, using samples in a particular neighbourhood around each grid point. In ordinary kriging, the constraint on the weights (see Appendix) results in the estimate being kept close to the neighbourhood data mean. The smaller the neighbourhood, the more constrained the estimate to the local means, and therefore the more local transitions on the kriged map. The risk of using a neighbourhood that is too small is the overall bias, the overall kriged mean differing from the data mean. Different neighbourhoods (disc radius and number of sample points) were tried and an overall non-bias criterion calculated. The neighbourhood retained was the one that minimized the quantity $\sum_{j=0}^{P}\left|\bar{z}_j - \bar{z}_j^{\mathrm{CK}}\right|$, where $\bar{z}_j$ is the average over sample locations, and $\bar{z}_j^{\mathrm{CK}}$ the average over the gridded kriged values (Figure 7). The grid mesh size was 15 miles in

latitude and 10 miles in longitude, compatible with the spatial resolution of the sampling. The neighbourhood was a disc of radius 25 miles. The minimum and maximum numbers of samples retained in the neighbourhood were 3 and 10, respectively. Some grid points were not estimated because of a lack of samples in their close neighbourhood.

The map of the anchovy length-frequency histogram was then derived from the maps of the polynomial coefficients using Equation (2), which provided a length frequency vector $y(x,l)$ for any length l, $l \in [l_1, l_2]$, at each grid node x.

3.5. Histogram maps in areas of probable fish presence

The anchovy length-frequency map was combined with the anchovy presence map. Only grid points where the probability of anchovy presence was >0.20 were considered. In areas where the probability was less, the observations were not considered consistent enough. The map shows the relative proportions of length classes in areas where anchovy were found consistently during the cruise series (Figure 8). Although frequency at length was estimated as a continuous curve, together with the estimation of variance at length, only selected length classes are shown in Figure 8, for the sake of clarity. There was consistently a greater proportion of small anchovy in areas off the mouths of the rivers Gironde (45.5°N 01.5°W) and Loire (47°N, 03°W). Large anchovy were proportionally more common offshore, the biggest concentrations in an area near the shelf break, south of 46°N, and along the 2°W meridian. Anchovy of medium size were found in most areas.

To illustrate the spatially explicit estimation of continuous length-frequency histograms, we selected three estimated histograms in three areas where large, small, and medium-sized fish were respectively dominant (Figure 9). It is of note that, owing to the polynomial decomposition, the estimated histogram at any location is a continuous curve, with possibly a complex, multimodal shape. The estimation variance can be high for specific length classes as a result of the combination of the squared polynomial values at length and the co-kriging variances (see Appendix).

Maps of the proportions of fish of length less or greater than a threshold were derived from Equation (3) in the most probable anchovy habitats (probability of presence >0.2; Figure 10). Large anchovy (length >16 cm) were dominant (relative proportion >0.6) in southern offshore habitats. In contrast, small anchovy (length <12 cm) were generally less dominant (relative proportions always <0.55), meaning that they consistently shared their habitats with large and medium-sized fish, in a given year and across years, even in the core areas of their spatial distribution.

4. Discussion

The procedure of Nerini *et al*. (2010) was well suited to mapping fish length-frequency histograms as spatially connected curves. The procedure is based on (i) a polynomial expansion, and (ii) co-kriging. The length-frequency histogram is modelled as a continuous curve using a polynomial expansion at each point, and co-kriging ensures coherence between the proportions of different length classes. Other methods can be applied to modelling the length-frequency histogram, but the mapping will necessarily require co-kriging. The reason for this is that co-kriging ensures that the probability of a

fish being in a given length interval is estimated coherently, so corresponds to the difference in the co-kriging estimates of the frequencies of the two bounds of the interval, which would not be the case if each bound was estimated by monovariate kriging. As a result, the estimated length frequencies naturally sum to unity, as do the length-frequency data.

The experimental histogram at each sample point could be discretized into *n* length groups corresponding to particular life stages, and the *n* frequency values could be co-kriged. This would allow for mapping of a discretized histogram. Here, the length-frequency histogram was modelled as a continuous curve, allowing for the mapping of any length value. Modelling curves using polynomial expansions is a classical procedure in functional data analysis (Ramsay and Silverman, 2005). Other curve-modelling methods were not attempted here. Another advantage of using orthogonal polynomials in modelling the histogram of fish length is that the polynomial coefficients can be simulated and combined geostatistically, facilitating simulation of the spatial distribution of the fish-length histogram.

If the curve to be modelled was known, a polynomial expansion could be fitted to match that curve exactly. Here, the true curve was unknown, and we selected a particular error level (the residual sum of squares) to estimate it. The order of the polynomial expansion influences the complexity and smoothness of the curves modelled. The residual variability around the modelled histograms was considered to be pure noise, and was not accounted for explicitly in the analysis. The residuals implicitly influenced the direct and cross-variograms, so were implicitly accounted for in the analysis. Polynomial expansions can be fitted to experimental histograms by regression or quadratic methods, because the polynomials form an orthogonal basis. In our case, polynomial coefficients were estimated by least squares. This was possible because the fish were measured with the same minimum resolution (0.5 cm) and all length classes were valued at each sample point. When experimental values are collected along $y(i,t)$ curves with different sampling resolutions, the curves sampled must be interpolated first along t: $\tilde{y}(i,t)$. Then the polynomial coefficients can be estimated quadratically as $z_k(i) = \int_\tau \tilde{y}(i,t)\, \phi_k(t)\, \mathrm{d}t$. Nerini *et al*. (2010) used that procedure. In our case, spline smoothing of the experimental length-frequency histograms generated occasional undesirable border effects, leading us to fit the polynomial coefficients by least squares.

Here, to select the number of polynomials, we used gof criteria based on the residual sum of squares. The Akaike information criterion (AIC: Chapter 2 of Burnham and Anderson, 2002) could also be used to choose between models with a varying number of parameters (polynomials), because it penalizes the residual statistics with the number of parameters. In our case, assuming Gaussian residuals, AIC suggested use of a similar number of polynomials (between 11 and 13: not shown) as the gof.

In our analysis we combined a map of the probability of fish presence with a map of fish length frequency, to characterize size-dependent anchovy habitats in spring in areas where anchovy was common in the surveys from 2000. Large anchovy specifically occupied highly probable habitats near the shelf break. In contrast, the highly probable coastal habitat off the Gironde estuary contained mainly small anchovy and larger length groups. Less-probable habitats north of 46.5°N off the Loire estuary held a mixture of small and medium-sized fish. These findings are consistent with previous knowledge of

the anchovy life cycle, derived from analysis of mean fish length (Chapter 8 of ICES, 2010; Petitgas *et al*., 2003b). However, the maps presented here are of better resolution and overall characterize the distribution of anchovy length groups more effectively. Such maps are also directly applicable for fisheries spatial management.

Further, the combination of fish length-frequency maps with maps of fish acoustic density opens the way for a fully spatially explicit approach to computing fish-stock acoustic abundance estimates at length, along with their estimation variance. Also, using orthogonal polynomials in modelling the histogram of fish length facilitates simulation of the spatial distribution of the fish-length histogram, and hence that of fish abundance at length.

As co-kriging preserves the functional relationships between variables in the estimates, the methodology is well suited to mapping the vectors of functionally related values characterizing fish communities, such as species proportions, length distributions, and growth curves. Our results therefore apply new tools, to extend the application of kriging techniques to curves or functions, opening new perspectives for mapping more complex information for ecosystem, conservation, or biodiversity studies.

Supplementary material

A simple average of the length distribution in the cells of the interpolation grid showing the advantages of co-kriging is provided as Supplementary material at the ICESJMS online version of this paper.

Acknowledgements

We thank the crews of the RV “Thalassa” for their support during the PELGAS surveys, and Pierre Beillois for maintaining the database. The study was partly supported by the project Reproduce of the MariFish EraNET. We thank the anonymous referees and editor for helping us improve the presentation.

References

Botsford, L., Brumbaugh, D., Grimes, C., Kellner, J., Largier, J., O’Farrell, M., Ralston, S., *et al*. 2009. Connectivity, sustainability, and yield: bridging the gap between conventional fisheries management and marine protected areas. Reviews in Fish Biology and Fisheries, 19: 69–95.

Burnham, K., and Anderson, D. 2002. Model Selection and Multimodel Inference, 2nd edn. Springer, New York.

Chilès, J-P., and Delfiner, P. 1999. Geostatistics: Modelling Spatial Uncertainty. John Wiley, New York.

Christensen, A., Mosegaard, H., and Jensen, H. 2009. Spatially resolved fish population analysis for designing MPAs: influence on inside and neighbouring habitats. ICES Journal of Marine Science, 66: 56–63.

Gautschi, W. 2004. Orthogonal Polynomials: Computation and Approximation. Oxford University Press, New York.

Goulard, M., and Voltz, M. 1992. Linear coregionalization model: tools for estimating and choice of multivariate variograms. Mathematical Geology, 24: 269–286.
ICES. 2010. Life cycle patterns of small pelagic fish in the Northeast Atlantic. ICES Cooperative Research Report, 306.
Nerini, D., Monestiez, P., and Manté, C. 2010. Cokriging for spatial functional data. Journal of Multivariate Analysis, 101: 409–418.
Petitgas, P., Massé, J., Beillois, P., Lebarbier, E., and Le Cann, A. 2003a. Sampling variance of species identification in fisheries acoustic surveys based on automated procedures associating acoustic images and trawl hauls. ICES Journal of Marine Science, 60: 437–445.
Petitgas, P., Massé, J., Grellier, P., and Beillois, P. 2003b. Variation in the spatial distribution of fish length: a multi-annual geostatistics approach on anchovy in Biscay, 1983–2002. ICES Document CM 2003/Q: 15.
Ramsay, J., and Silverman, B. 2005. Functional Data Analysis, 2nd edn. Springer, New York.
Renard, D., and Bez, N. 2008. RGeoS: Geostatistical package. http://cg.ensmp.fr/rgeos/.
Rivoirard, J., Simmonds, J., Foote, K., Fernandes, P., and Bez, N. 2001. Geostatistics for Estimating Fish Abundance. Blackwell Science Ltd, London.
van Keeken, O., van Hoppe, M., Grift, R., and Rijnsdorp, A. 2007. Changes in the spatial distribution of North Sea plaice (*Pleuronectes platessa*) and implications for fisheries management. Journal of Sea Research, 57: 187–197.
Wackernagel, H. 1995. Multivariate Geostatistics: an Introduction with Applications. Springer, Berlin.
Werner, E. E., and Gilliam, J. F. 1984. The ontogenic niche and species interactions in size-structured populations. Annual Review of Ecology and Systematics, 15: 393–425.
Woillez, M., Rivoirard, J., and Fernandes, P. G. 2009. Evaluating the uncertainty of abundance estimates from acoustic surveys using geostatistical simulations. ICES Journal of Marine Science, 66: 1377–1383.

Appendix

Kriging and co-kriging equations

Kriging is a linear estimator that is by construction unbiased and of minimum variance. Although equations can be found in geostatistical text books (e.g. Wackernagel, 1995; Chilès and Delfiner, 1999), we here provide with the same notations and in matrix form the kriging and co-kriging equations necessary for the purpose of the current study.

Ordinary kriging with moving neighbourhood of the indicator of fish presence

The target variable is the indicator of fish presence. The mean of the target variable is unknown, but assumed to be constant or varying sufficiently smoothly to be considered as constant in a local neighbourhood around the grid points. The kriged estimate of the indicator of fish presence at point x_0 is $1^K(0) = \sum_{\alpha \in V_0} \lambda(\alpha) 1(\alpha)$, where α is the index of sample point x_α in the vicinity V_0 of x_0. The estimate is unbiased if the weights λ satisfy

$\sum_{\alpha \in V_0} \lambda(\alpha) = 1$. The kriging variance is the minimum estimation variance, and is obtained from the weights that solve the kriging system:

$$\begin{bmatrix} \Gamma_{\alpha\beta} & I \\ I' & 0 \end{bmatrix} \begin{bmatrix} \Lambda_\beta \\ \mu \end{bmatrix} = \begin{bmatrix} \Gamma_{\alpha 0} \\ 1 \end{bmatrix},$$

where: $\Gamma_{\alpha\beta}$ is the matrix block of dimension (n_0,n_0), where entry $\gamma(\alpha,\beta)$ is the variogram value for the distance $|x_\alpha - x_\beta|$ between the n_0 samples in the neighbourhood V_0 ; $\Gamma_{\alpha 0}$ is the column vector of dimension n_0 where entry $\gamma(\alpha,0)$ is the variogram value for the distance $|x_\alpha - x_0|$ between sample x_α of the neighbourhood V_0 and grid node x_0; Λ_β is the column vector of dimension n_0 where entry λ_β is the kriging weight assigned to sample x_β of the neighbourhood V_0; and I is a column vector of unit values of dimension n_0 and μ a Lagrange multiplier.

The kriging variance is $\sigma_K^2 = \Lambda_\beta' \Gamma_{\alpha 0} + \mu$.

Ordinary co-kriging with moving neighbourhood of the polynomial coefficients

The target variables are each of the P+1 polynomial coefficients. The co-kriging estimate $z_j^{CK}(0)$ of the polynomial coefficient of degree j at grid node x_0 is a linear combination of all the polynomial coefficients (indexed by k) at stations x_α in neighbourhood V_0 of x_0:

$$z_j^{CK}(x0) = \sum_{k=0}^{P} \sum_{\alpha \in V0} \lambda_{kj}(\alpha) z_k(\alpha) \quad , j = \{0, 1, \ldots, P\}.$$

Unbiasedness ($E[z_j(x_0) - z_j^{CK}(x_0)] = 0$) is achieved by applying the following constraints on the weights:

$\sum_{\alpha \in V_0} \lambda_{kj}(\alpha) = \delta_{kj}$ where $\delta_{kj}=1$ if $j=k$ and 0 otherwise, $k=\{0, \ldots, P\}$, $j=\{0, \ldots, P\}$.

The weights that minimize the estimation variance $E[(z_j(x_0 - z_j^{CK}(x_0))^2]$ under the unbiasedness constraints are the solution of the following linear (co-kriging) system (Chapter 5 of Chilès and Delfiner, 1999):

$$\begin{bmatrix} \Gamma_{kk'} & F_{kk'} \\ F_{kk'}' & 0 \end{bmatrix} \begin{bmatrix} \lambda_{kj} \\ \mu_{kj} \end{bmatrix} = \begin{bmatrix} \Gamma_{kj} \\ \delta_{kj} \end{bmatrix}$$

where: $\Gamma_{kk'} = \begin{bmatrix} \gamma_{00}(\alpha,\beta) & \ldots & \gamma_{0k}(\alpha,\beta) & \ldots & \gamma_{0P}(\alpha,\beta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma_{k0}(\alpha,\beta) & \ldots & \gamma_{kk}(\alpha,\beta) & \ldots & \gamma_{kP}(\alpha,\beta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma_{P0}(\alpha,\beta) & \ldots & \gamma_{Pk}(\alpha,\beta) & \ldots & \gamma_{PP}(\alpha,\beta) \end{bmatrix}$ is the matrix block of dimension ($(P+1)n_0$,$(P+1)n_0$), where entry $\gamma_{kk'}(\alpha,\beta)$ is a block of dimension (n_0,n_0) containing the cross-variogram values between variables k and k' for all distances $|x_\alpha - x_\beta|$ between the

n_0 samples in the neighbourhood V_0 ; $\Gamma_{kj} = \begin{bmatrix} \gamma_{0j}(\alpha,0) \\ ... \\ \gamma_{kj}(\alpha,0) \\ ... \\ \gamma_{Pj}(\alpha,0) \end{bmatrix}$ is the column vector of dimension $(P+1)n_0$, where entry $\gamma_{kj}(\alpha,0)$ is a column vector of dimension n_0 containing the cross-variogram values between variable k and variable j for all distances $|x_\alpha - x_0|$ between samples in neighbourhood V_0 and the grid point x_0; $\lambda_{kj} = \begin{bmatrix} \lambda_{0j}(\beta) \\ ... \\ \lambda_{kj}(\beta) \\ ... \\ \lambda_{Pj}(\beta) \end{bmatrix}$ is the column vector of dimension $(P+1)n_0$, where entry $\lambda_{kj}(\beta)$ is the column vector of dimension n_0 containing the co-kriging weights assigned to variable k at sample locations x_β for estimating variable j at location x_0; $\mu_{kj} = \begin{bmatrix} \mu_{0j} \\ ... \\ \mu_{kj} \\ ... \\ \mu_{Pj} \end{bmatrix}$ is the column vector of dimension $(P+1)$ where entry μ_{kj} is the Lagrange parameter assigned to variable k for estimating variable j; $F_{kk'}$ is a matrix block of dimension $((P+1)n_0,(P+1))$ where the k^{th} column is a vector made of zero values except for the n_0 lines corresponding to variable k, where values are 1; $F_{kk'}^{'}$ is the transpose of $F_{kk'}$; 0 is a matrix block of dimension $((P+1),(P+1))$ filled with zero values; and δ_{kj} is a column vector of dimension $(P+1)$ corresponding to the constraints on the sum of the kriging weights when estimating variable j, where δ_{kj} equals 1 if $k = j$, and 0 otherwise.

Using the above notations, the co-kriging estimation variance for variable j is

$$\sigma_{\mathrm{CK}}{}^2(j) = \lambda_{kj}{}^{'}\,\Gamma_{kj} + \mu_{jj}\,.$$

Developing the estimation variance of the curve $y(0,t)$ requires computation of the covariance between estimation errors $E[e_{j1}e_{j2}]$ for pairs of variables j_1, j_2:

$$E[(y(0,t) - y(0,t)^{\mathrm{CK}})^2] = \sum_{j=0}^{P} \phi_j(t)^2 \sigma_{\mathrm{CK}}{}^2(j) + \sum_{j1 \neq j2} \sum \phi_{j1}(t)\,\phi_{j2}(t) E[e_{j1}e_{j2}]$$

$$E[e_{j1}e_{j2}] = \left[\lambda_{kj2}\right]^{'} \left[\Gamma_{kj1}\right] + \mu_{j2\,j1}\,.$$

Computation of the covariance between estimation errors requires the left-side term [λ_{kj} , μ_{kj}]' and the right-side term [Γ_{kj} , δ_{kj}]' of the co-kriging system. In this study, we did not consider the covariance terms between estimation errors, leading to a probable inflation of the estimation variance of $y(0,t)$.

Figure legends

Figure 1. Location of the midwater trawl hauls used in this study, all years pooled, 2000–2010.

Figure 2. Probability map of anchovy presence, 2000–2010.

Figure 3. Boxplots of the goodness-of-fit criterion (gof) when fitting the observed fish length-frequency histograms at sampled stations with Legendre polynomial expansions of increasing degree. The dashed line represents a gof value of 0.05.

Figure 4. Multivariate fit of the direct variograms, with each variogram modelled as the sum of a nugget effect and a spherical model with a range of 20 miles. The sills are fitted automatically using the algorithm of Goulard and Voltz (1992). The cross-variogram parameters are shown in Figure 5.

Figure 5. Multivariate fit of the cross-variograms. The figure plots the sills $b_u(i,j)$ (u = {1,2}) fitted to each of the two elementary structures for all cross-variograms i–j ($i<j$). Indices 1 and 2 identify the nugget (abscissa) and the spherical component (ordinate), respectively; i and j indices identify the variables: i = {0, …, P–1}, j = {1,…,P}, $i<j$.

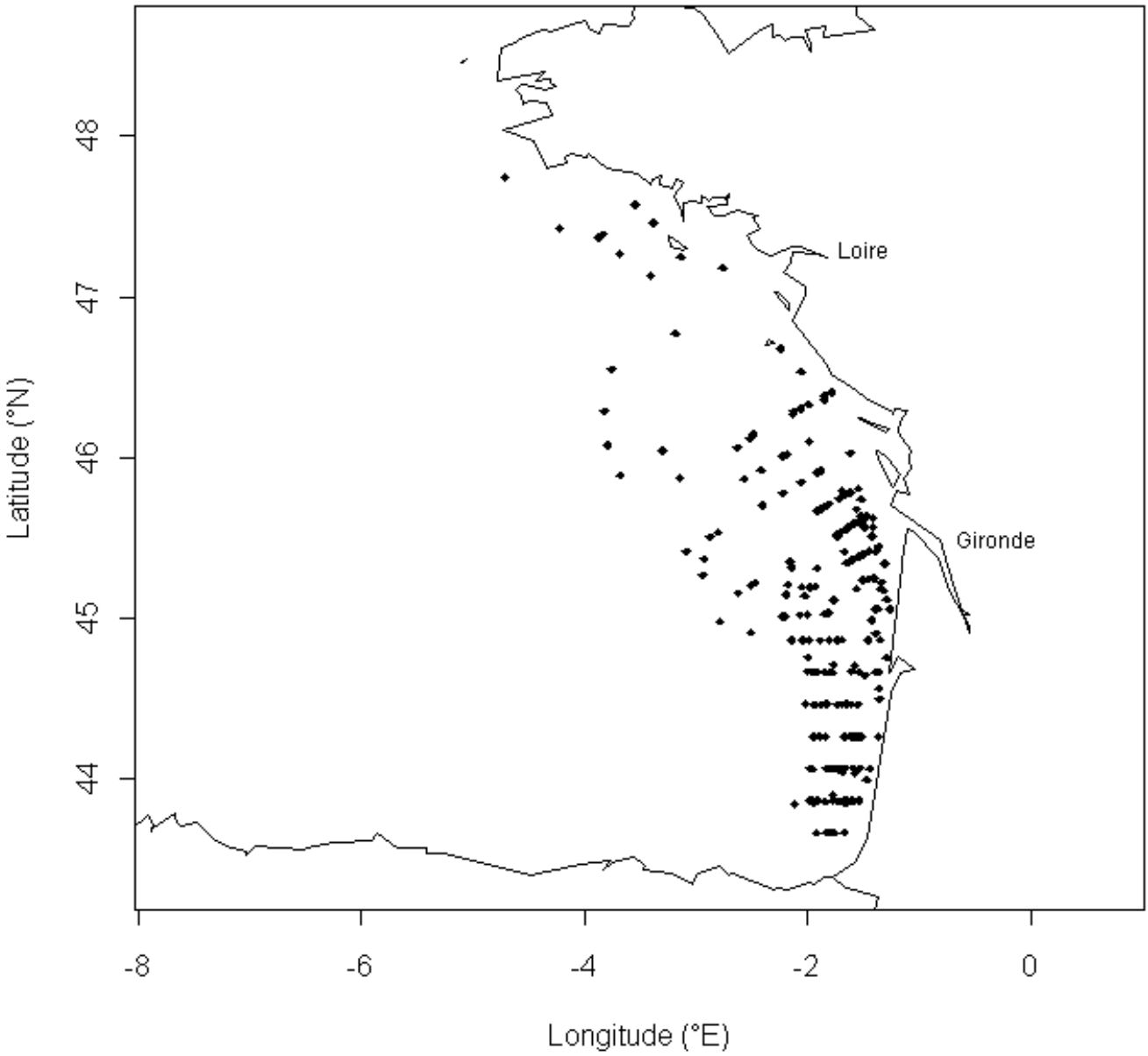
Figure 6. Four examples of fitted cross-variograms models selected from the plot in Figure 5. Upper left, cross-variogram between variables z_4 and z_6, with the two components (nugget and spherical) well represented; lower left, cross-variogram between variables z_6 and z_7, with the components little represented; upper right, cross-variogram between variables z_9 and z_{13}, with both components half represented; lower right, cross-variogram between variables z_4 and z_{11}, with the spherical component half represented and the nugget little represented. The black lines represent the limits of model acceptability, according to the positivity constraint.

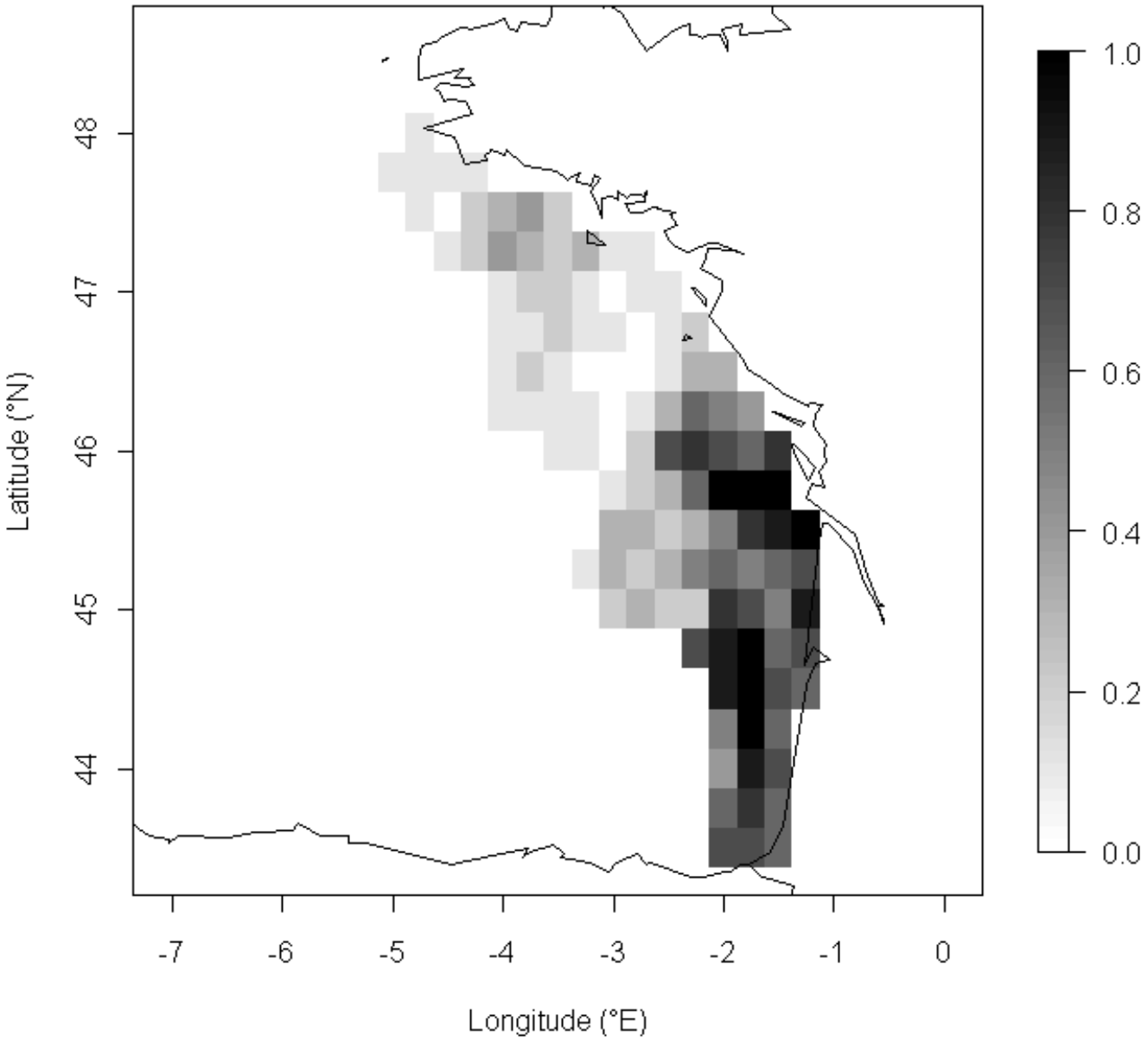
Figure 7. Boxplots of the polynomial coefficients: left, data; right, kriged maps.

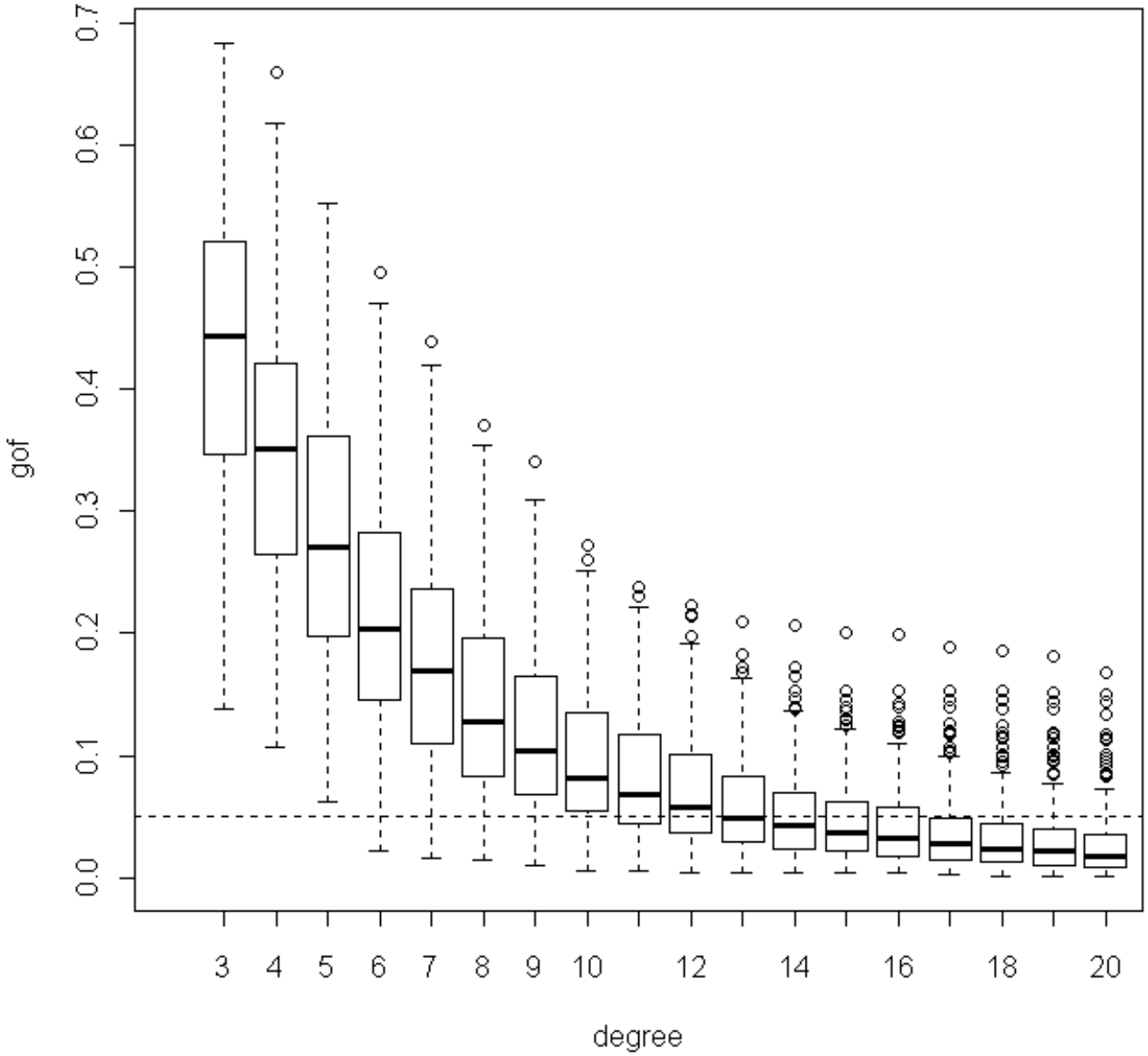
Figure 8. Map of anchovy length-frequency histograms in areas where the probability of anchovy presence is >0.2.

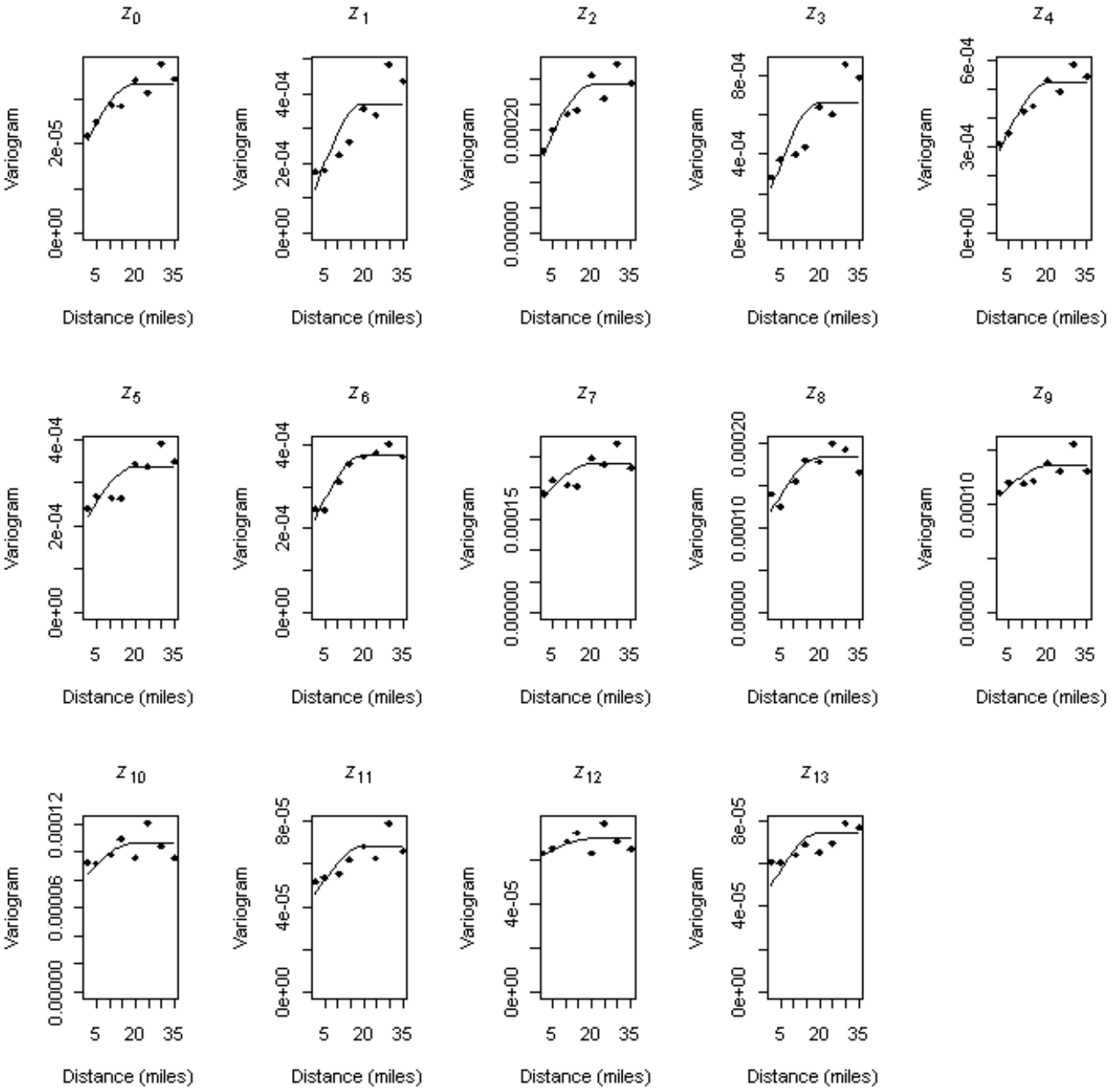
Figure 9. Examples of estimated anchovy length-frequency histograms at points numbered 135, 178, and 310. The vertical bars indicate a confidence interval of ±2 estimated standard deviation above zero. The map in the upper right corner shows the location where the kriged histograms are situated.

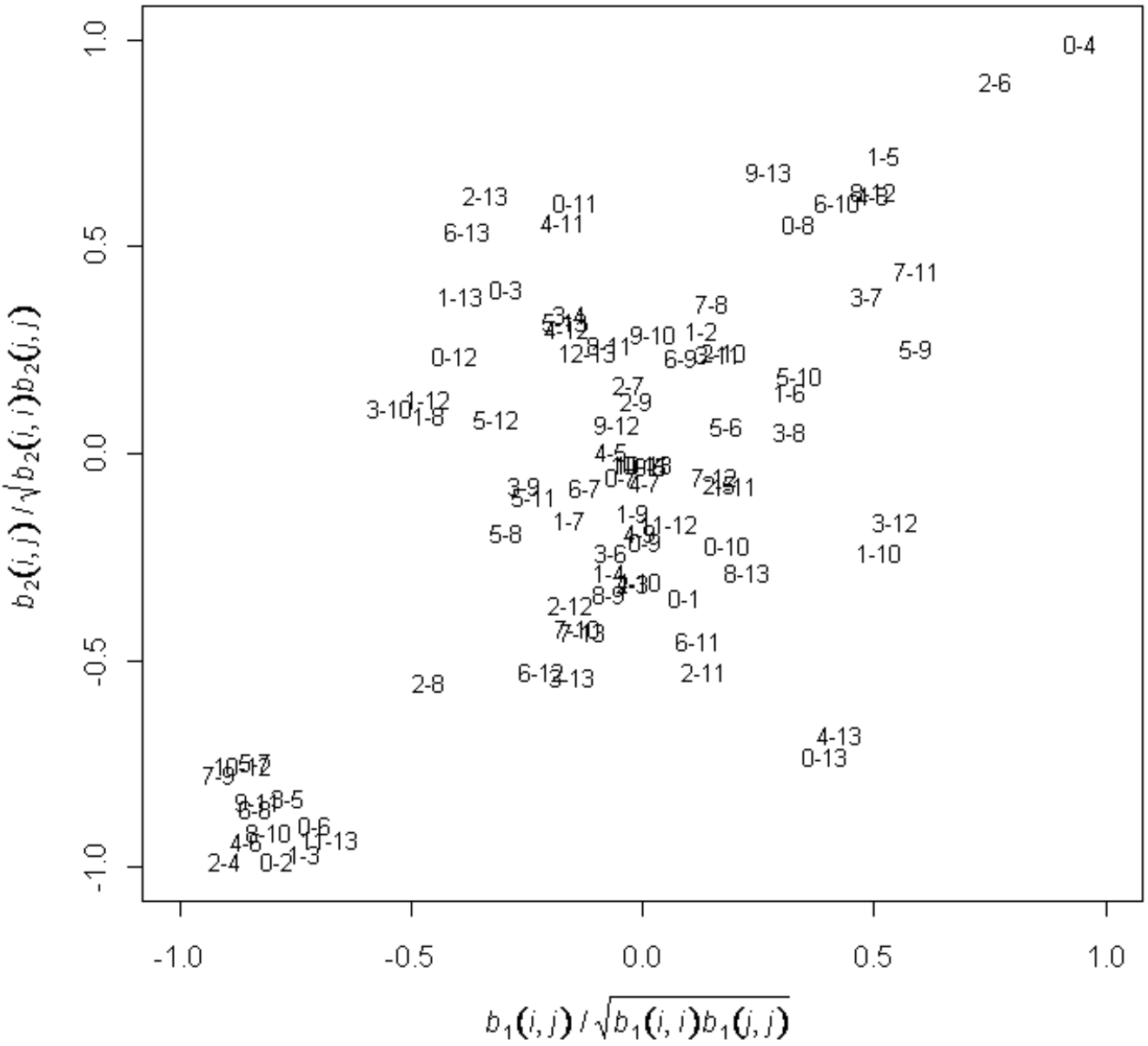
Figure 10. Maps of the proportion of anchovy <12 cm (left) and >16 cm (right) in areas where the probability of anchovy presence is >0.2.

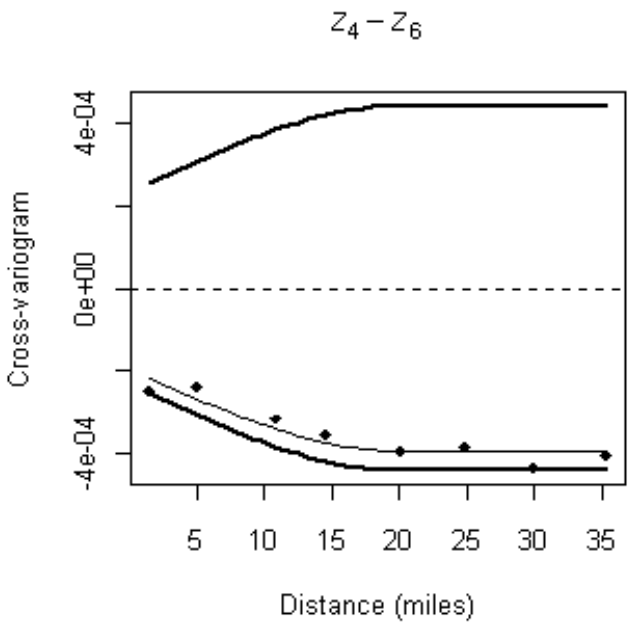

Loire
Gironde
48
47
46
45
44
Latitude (°N)
-8
-6
-4
-2
0
Longitude (°E)

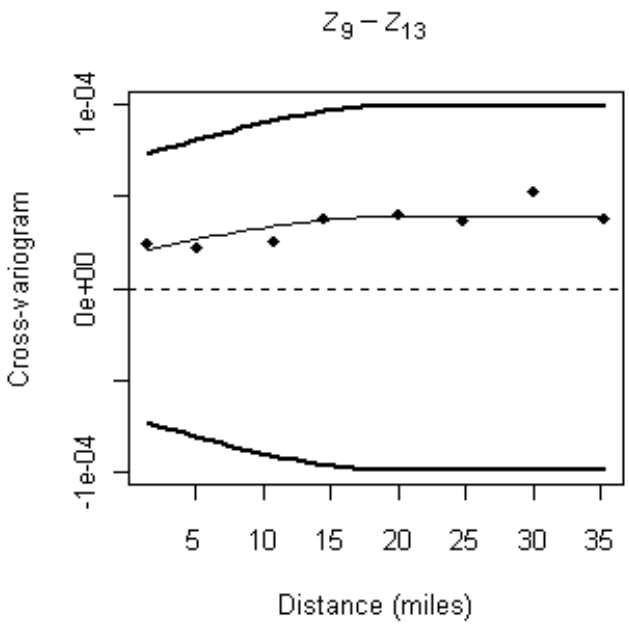

Latitude (°N)
48
47
46
45
44
-7
-6
-5
-4
-3
-2
-1
0
Longitude (°E)
1.0
0.8
0.6
0.4
0.2
0.0

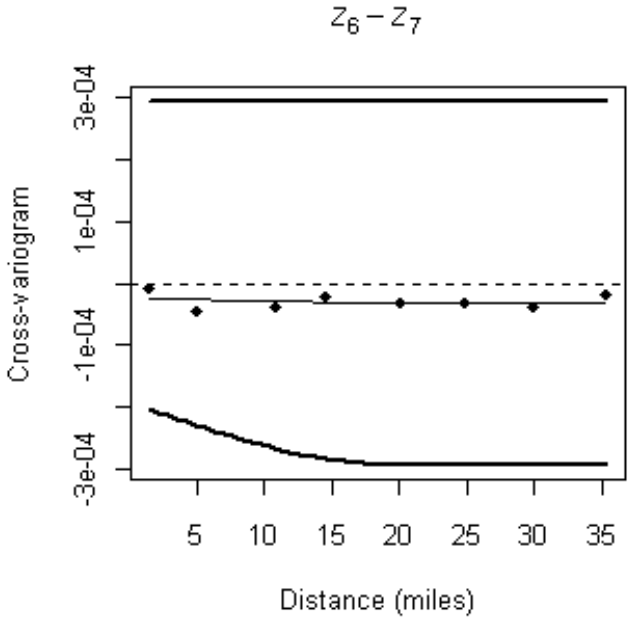

gof
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
3
4
5
6
7
8
9
10
12
14
16
18
20
degree

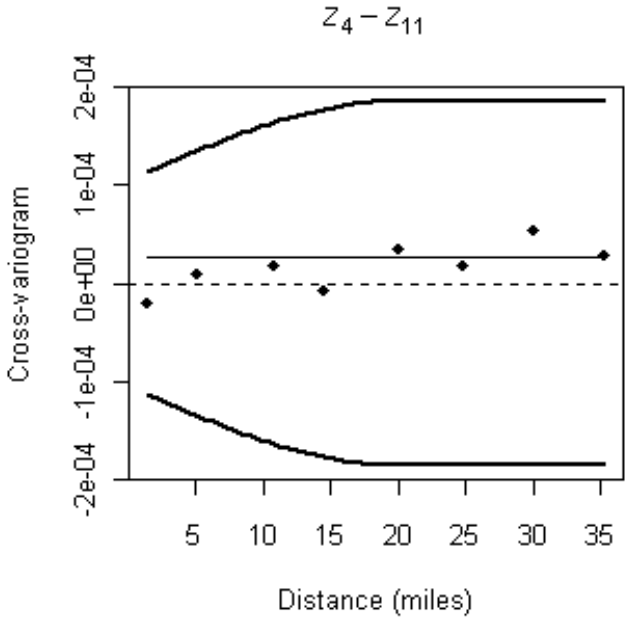

z_0
z_1
z_2
z_3
z_4
z_5
z_6
z_7
z_8
z_9
z_{10}
z_{11}
z_{12}
z_{13}
Variogram
Distance (miles)
0e+00
2e-05
4e-04
0.00020
6e-04
3e-04
0.00015
0.00010
0.00012
0.00006
8e-05
5
20
35

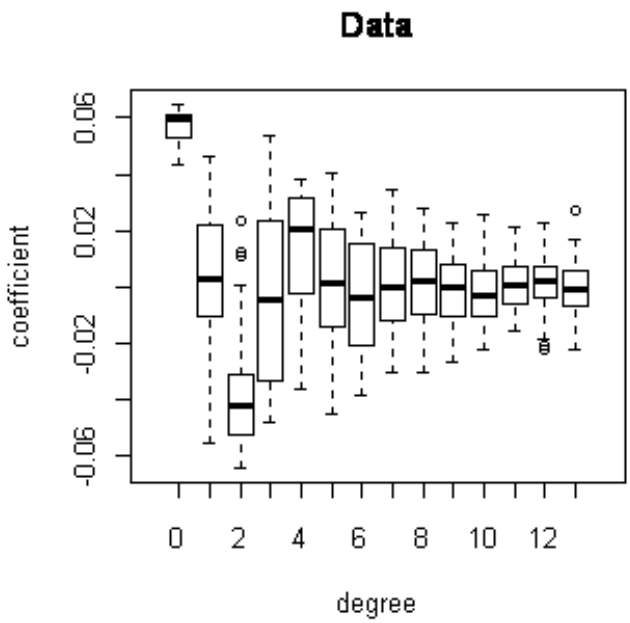

1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
$b_2(i,j)/\sqrt{b_2(i,i)b_2(j,j)}$
$b_1(i,j)/\sqrt{b_1(i,i)b_1(j,j)}$
0-4
2-6
1-5
9-13
2-13
0-11
6-10
6-13
4-11
0-8
7-11
1-13
0-3
7-8
3-7
9-10
1-2
0-12
6-9
5-9
5-10
1-6
2-7
2-9
3-10
1-8
5-12
9-12
5-6
3-8
3-9
6-7
1-9
1-7
5-8
3-12
0-10
1-10
3-6
8-13
0-1
6-11
2-11
2-8
4-13
0-13
2-4
0-2
1-3

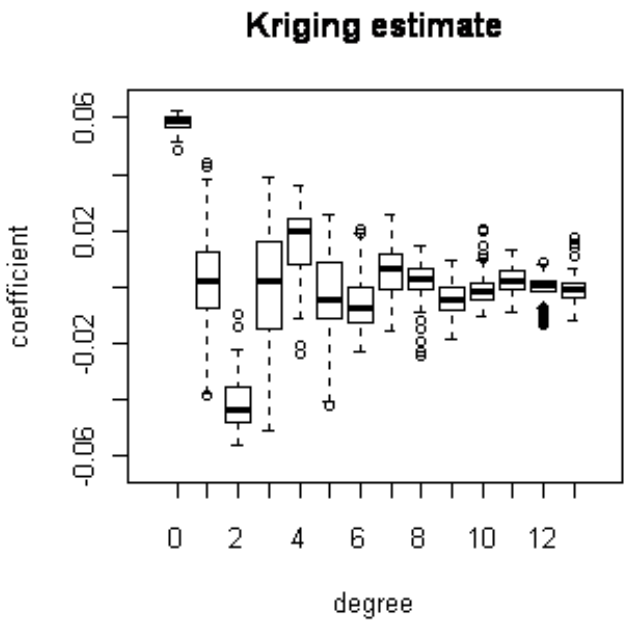

$Z_4 - Z_6$
Cross-variogram
4e-04
0e+00
-4e-04
5
10
15
20
25
30
35
Distance (miles)

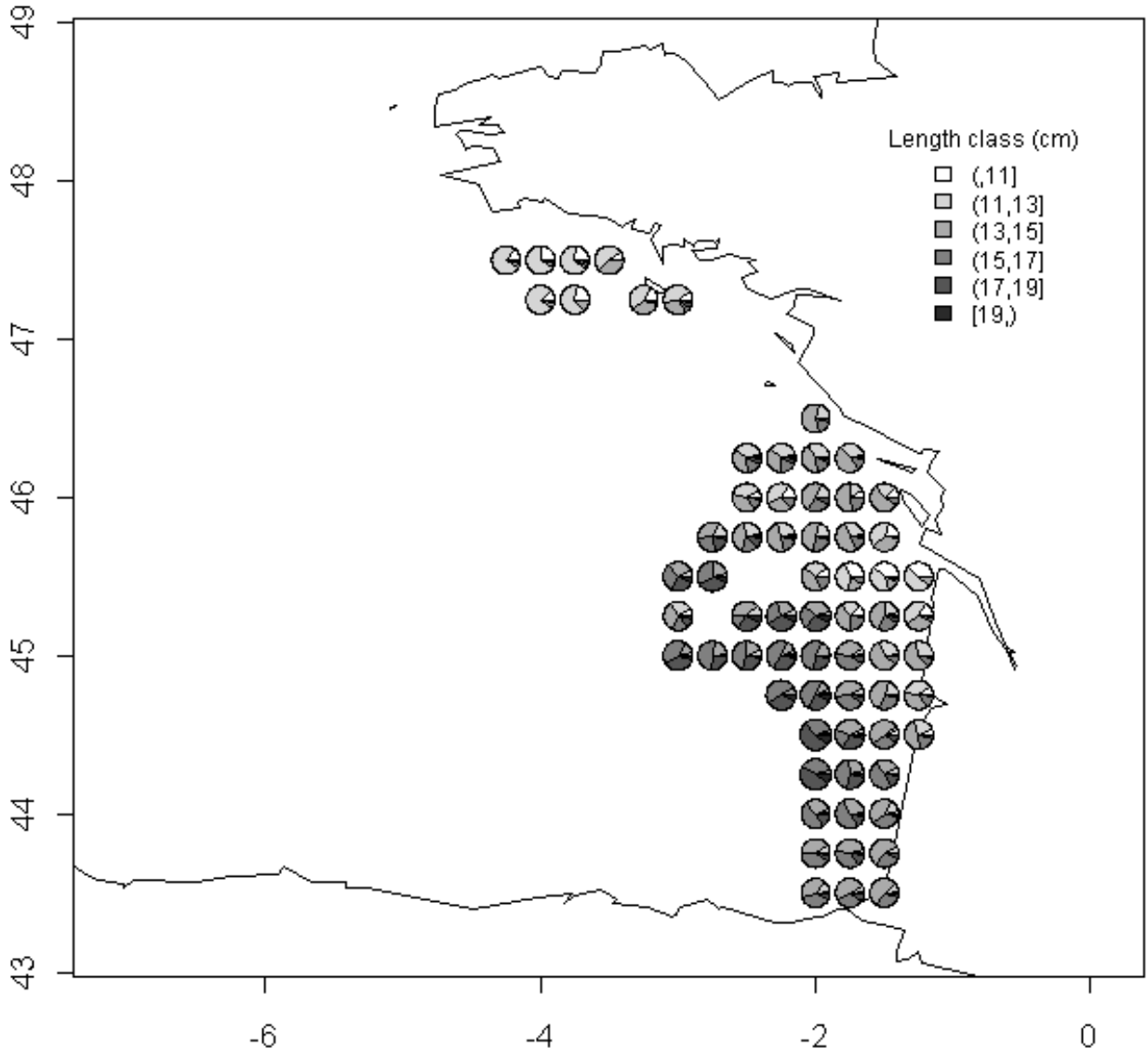

$Z_9 - Z_{13}$
Cross-variogram
1e-04
0e+00
-1e-04
5
10
15
20
25
30
35
Distance (miles)

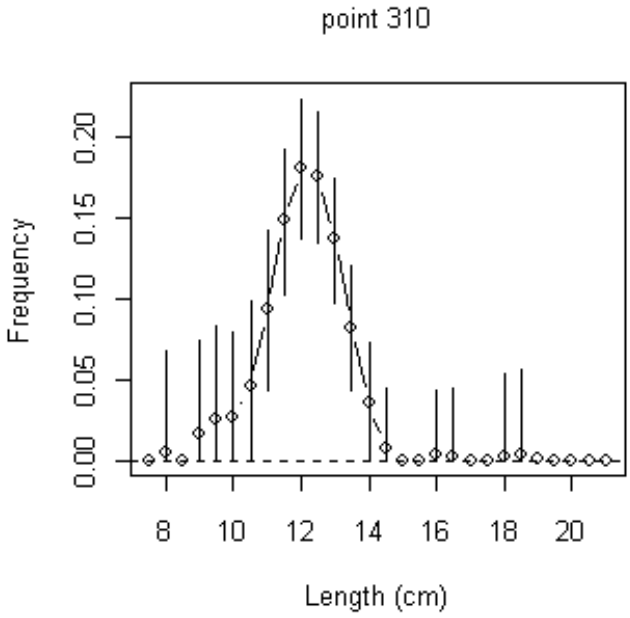

$Z_6 - Z_7$
Cross-variogram
3e-04
1e-04
-1e-04
-3e-04
5
10
15
20
25
30
35
Distance (miles)

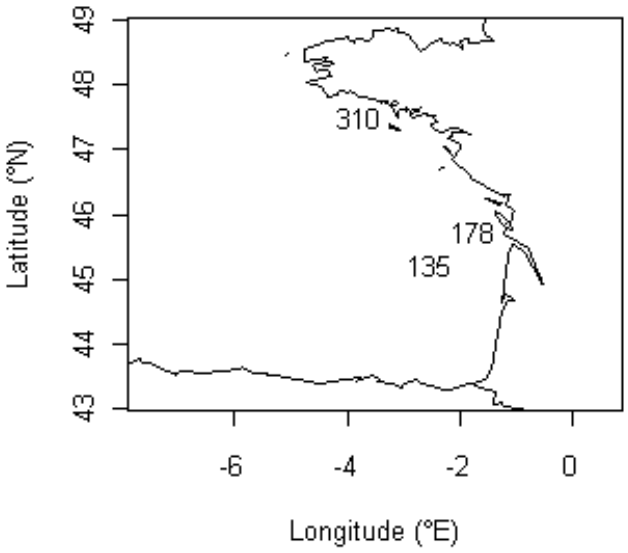

$Z_4 - Z_{11}$
Cross-variogram
2e-04
1e-04
0e+00
-1e-04
-2e-04
5
10
15
20
25
30
35
Distance (miles)

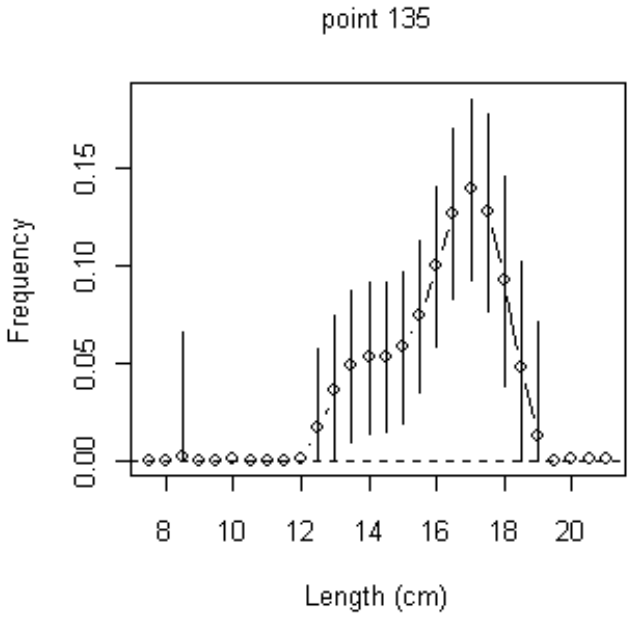

Data
coefficient
0.06
0.02
-0.02
-0.06
0
2
4
6
8
10
12
degree

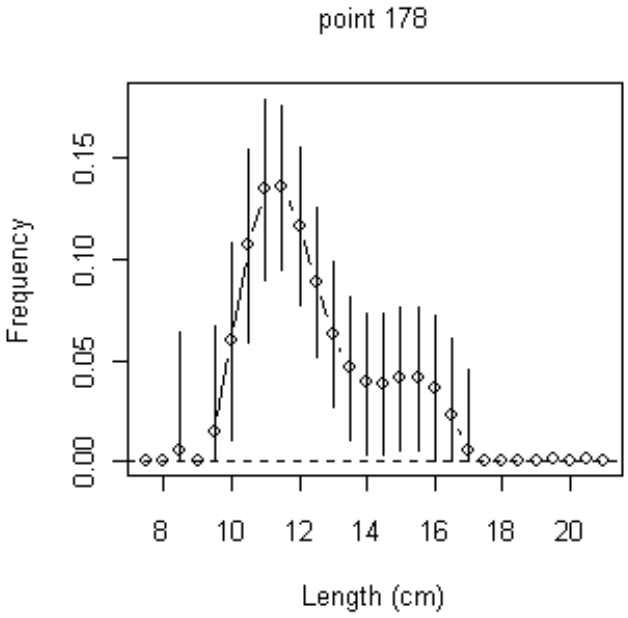

Kriging estimate
coefficient
0.06
0.02
-0.02
-0.06
0
2
4
6
8
10
12
degree

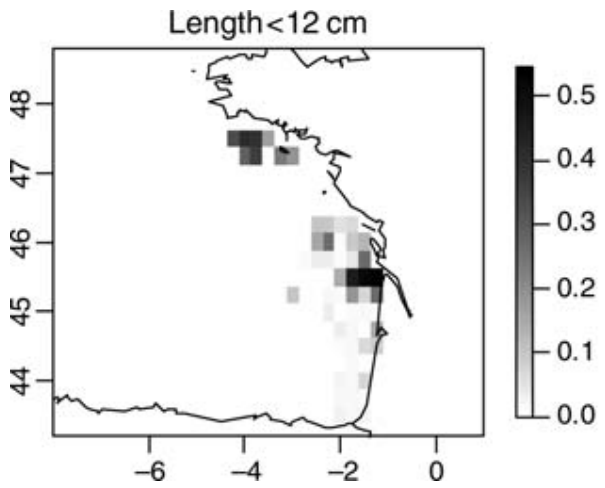

Length class (cm)
(,11]
(11,13]
(13,15]
(15,17]
(17,19]
[19,)
49
48
47
46
45
44
43
-6
-4
-2
0

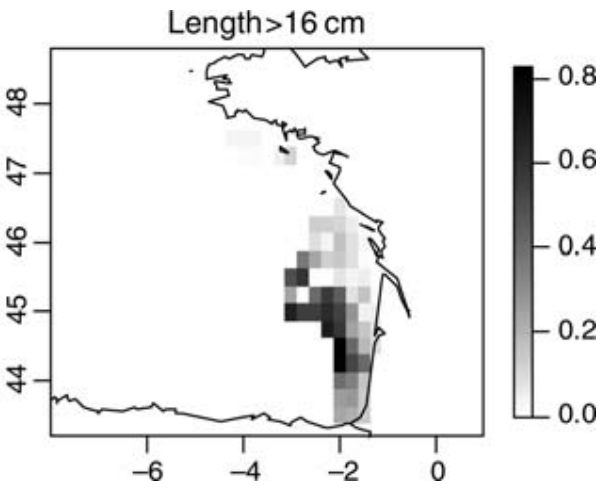

point 310
Frequency
0.00
0.05
0.10
0.15
0.20
8
10
12
14
16
18
20
Length (cm)


Latitude (°N)
43
44
45
46
47
48
49
310
178
135
-6
-4
-2
0
Longitude (°E)


point 135
Frequency
0.00
0.05
0.10
0.15
8
10
12
14
16
18
20
Length (cm)


point 178
Frequency
0.00
0.05
0.10
0.15
8
10
12
14
16
18
20
Length (cm)


Length<12 cm
48
47
46
45
44
−6
−4
−2
0
0.5
0.4
0.3
0.2
0.1
0.0


Length>16 cm
48
47
46
45
44
−6
−4
−2
0
0.8
0.6
0.4
0.2
0.0

Supplementary material: ICES Journal of Marine Science, 68

Spatially explicit estimation of fish-length histograms, with application to anchovy habitats in the Bay of Biscay

Pierre Petitgas, Mathieu Doray, Jacques Massé, and Patrick Grellier

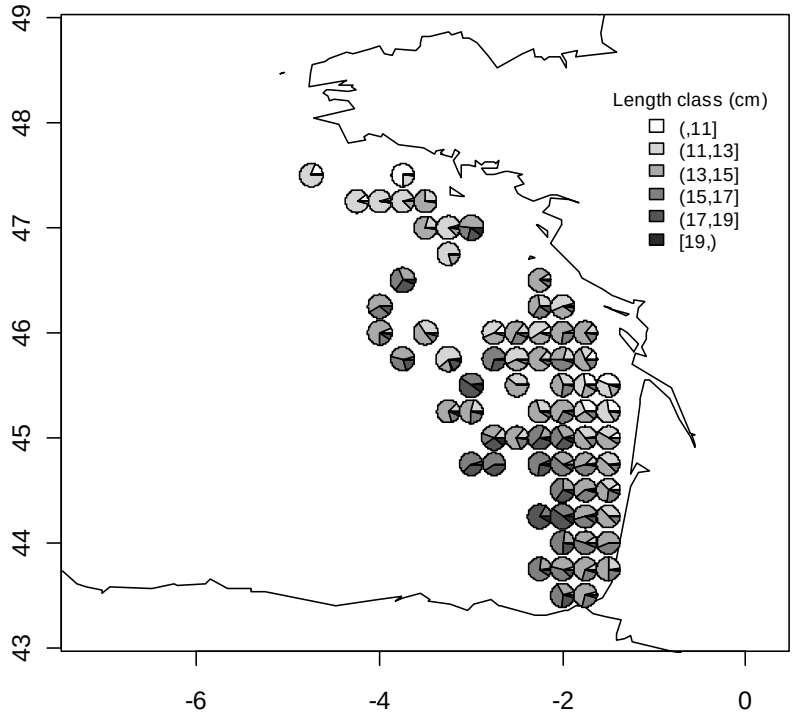


Figure S1. The simple average of the length distribution in the cells of the interpolation grid. In comparison, mapping the histogram by co-kriging has the advantages of (i) optimal interpolation at each point of the interpolation grid, and (ii) calculation of the corresponding estimation variance.