A Geostatistical Definition of Hotspots for Fish Spatial Distributions

Petitgas Pierre [1, *], Woillez Mathieu [2], Doray Mathieu [1], Rivoirard Jacques [3]

[1] IFREMER, Research Unit EMH, Rue de l'Île d'Yeu, 44300 Nantes, France
[2] IFREMER, Research Unit STH, Pointe du Diable, 29280 Plouzané, France
[3] MINES ParisTech, Centre for Geosciences and Geoengineering, PSL Research University, 35 rue Saint Honoré, 77305 Fontainebleau, France

* Corresponding author : Pierre Petitgas, email address : pierre.petitgas@ifremer.fr

Abstract :

Research surveys at sea are undertaken yearly to monitor the distribution and abundance of fish stocks. In the survey data, a small number of high fish concentration values are often encountered, which denote hotspots of interest. But statistically, they are responsible for important uncertainty in the estimation. Thus understanding their spatial predictability given their surroundings is expected to reduce such uncertainty. Indicator variograms and cross-variograms allow to understand the spatial relationship between values above a cutoff and the rest of the distribution under that cutoff. Using these tools, a "top" cutoff can be evidenced above which values are spatially uncorrelated with their lower surroundings. Spatially, the geometric set corresponding to the top cutoff corresponds to biological hotspots, inside which high concentrations are contained. The hotspot areas were mapped using a multivariate kriging model, considering indicators in different years as covariates. The case study considered here is the series of acoustic surveys Pelgas performed in the Bay of Biscay to estimate anchovy and other pelagic fish species. The data represent tonnes of fish by square nautical mile along transects regularly spaced. Top cutoffs were estimated in each year. The areas of such anchovy hotspots are then mapped by co-kriging using all information across the time series. The geostatistical tools were adapted for estimating hotspot habitat maps and their variability, which are key information for the spatial management of fish stocks. Tools used here are generic and will apply in many engineering fields where predicting high concentration values spatially is of interest.

Keywords : Hotspots, Indicators, Non-linear geostatistics, Multivariate geostatistics, Anchovy, Biscay

1 Introduction

In fisheries ecology the concept of hotspots refers to areas of high biological concentration due to particular environmental conditions shaping locally ecosystem productivity and structure. The concept of hotspots has been applied to single species (fishing hotspots: Yasuda et al. 2014) or species richness (biodiversity hotspots: Myers et al. 2000; Stuart-Smith et al. 2013) as well as areas of enhanced trophic interactions (Santora et al. 2012). The identification of hotspots is a fundamental aspect of conservation and spatial management strategies. Fisheries management now includes more global conservation issues and the approach developed here to define hotspots is part of such global strategy. Nelson and Boots (2008) review methods for detecting hotspots of over abundance. Typically, a threshold is applied to differentiate between hot and non-hot areas and this threshold can be defined globally based on the cumulative distribution of survey data or locally using a kernel. It is more usual in fisheries ecology to define a threshold globally. Defining a threshold is somewhat subjective and attempts have been made to diminish subjectivity by defining thresholds based on rules. Bartolino et al. (2011) suggest a rule for a global approach that is based on the shape of the cumulative distribution of the data. Here a local rule to define hotspots is suggested that is based on a non-linear geostatistical approach. Our definition of hotspots is based on the spatial relationship between values above a cutoff and the rest of the distribution under that cutoff. It results in defining a "top" cutoff which translates spatially in geometric sets that can be mapped by kriging. The advantages of doing so are that the cutoff defining hotspots is based on local transition probabilities, will vary across years with global abundance and does not need to be fixed. In effect, the notion of high value is relative and varies between years depending on global fish abundance. The fisheries survey data series

analyzed spanned fourteen years and showed variation of maximum values and global abundance across years, requiring a generic approach to defining hotspots. The resource is the anchovy in the Bay of Biscay. The fishery has been closed in the period 2005 to 2010 and spatial management measures have been tested (Lehuta et al. 2010). In addition to suggesting a methodology for defining hotspots, the present study also provides insight into the uncertainty when mapping hotspot habitats due to interannual variability. First, it is shown how the variograms and cross-variograms of indicators (Rivoirard 1994) allow to analyze the spatial organization of geometric sets defined by a series of indicators and how these properties can be used to define the cutoff for hotspots. The series of cutoffs along years results in a series of indicators. Then each indicator in each year is considered as a covariate and a linear model of coregionalization (Chilès and Delfiner 2012) is used to estimate the hotspots in each year given those in all years. The procedure is then applied on the survey series of anchovy distribution in the Bay of Biscay. The procedure is compared to simple thresholding to further demonstrate its appropriateness.

2 Methods

2.1 A Geostatistical definition of biological hotspots

Let $Z(x)$ be a random function taking values at points x (which will represent anchovy concentration for a given year). The indicator of the geometrical set A_i defined by cutoff z_i is

$$1_{Ai}(x) = \{1 \, if \, x \in Ai, Z(x) \geq zi\}$$
$$\{0 \, if \, x \notin Ai, Z(x) < zi\} .$$

The variogram of the indicator of A_i measures the probability that a vector of distance h has one extremity inside and the other outside the set A_i

$$\gamma_i(h) = P[Z(x) \in Ai, Z(x+h) \notin Ai] .$$

It is the probability of entering into A_i when going from point x to point x+h.

Let $z_j > z_i$ be another cutoff. The cross-variogram of the indicators of A_j and A_i measures the probability that a vector of distance h has one extremity inside A_j and the other outside A_i

$$\gamma_{ixj}(h)=P[Z(x)\in Aj, Z(x+h)\notin Ai].$$

It is the probability of going from outside A_i at one point to inside A_j at another point h apart from the first one.

The ratio of $\gamma_{ixj}(h)$ on $\gamma_i(h)$ measures the transition probability of getting inside A_j when entering into A_i from distance h

$$\frac{\gamma_{ixj}(h)}{\gamma_i(h)}=P[Z(x)\in Aj/Z(x)\in Ai, Z(x+h)\notin Ai].$$

An increasing ratio with distance h means that the sets A_j tend to be positioned in the middle of the sets A_i. In contrast, if the variogram ratio does not vary with distance (i.e., stays flat), the sets A_j are positioned at random within the sets A_i and thus, being in the set A_i does not inform on where sets A_j are located (no edge effect within A_i: Fig. 1). Values greater than z_j are spatially uncorrelated with the geometry of A_i. The spatial behavior within A_i of the indicator of values greater than z_j, is taken in charge by the residuals of the regression of

$1\{Z(x)\geq zj\}$ on $1\{Z(x)\geq zi\}$ (model with orthogonal residuals: Rivoirard 1994). If the cutoff z_j is sufficiently high, due to the destructuration of high values (Matheron 1982), the residual will often be spatially unstructured (pure nugget effect). The question is then to find a high enough cutoff z_i for which A_j (j > i) has no edge effect within A_i and possibly above which the residuals are pure nugget.

Based on this property, it is suggested to consider this set A_i as representing the hotspots of the year. Higher values are spatially uncorrelated with the geometry of A_i and cannot be predicted within A_i. The cutoff z_i plays the role of the topcut in the model proposed by Rivoirard et al. (2013). In practice, z_i can be estimated using the following procedure. Consider a suite of increasing cutoffs ($z_1<z_2<\ldots z_i<z_j\ldots z_n$). The hotspot habitat of the year is

defined as the geometrical set for cutoff z_i and is noted Ai, where i in {1,2,…(n-1)} is the

minimum rank for which the ratio of $\gamma_{ix(i+1)}(h)$ over $\gamma_i(h)$ is flat (the ratio does not vary

with distance h) and possibly for which the variogram of the residuals of $1\{Z(x) \geq z_{(i+1)}\}$ on

$1\{Z(x) \geq z_i\}$ is also flat.

2.2 Multi-year co-kriging of hotpsots

For each year, hotspots are defined by the indicator of one A_i. The cutoff rank *i* will vary

among years. So applying the previous procedure on a time series of P data sets results in P

indicators of hotspots. It is expected that the hotspots will show spatial correlation over the

time series and therefore the P indicators are expected to be spatially correlated. To map

hotspots in any given year k in{1,… P} the indicator in each year was co-kriged with all P

indicators from all years. As all P indicators are defined on the same grid, co-kriging is

homotopic and each indicator in each year was estimated by co-kriging with the other

indicators from all years (Chilès and Delfiner 2012). For that a linear model of

coregionalization was fitted using the automated fitting procedure of Dessasis and Renard

(2013). All computations were done with the RGeostats package (Renard et al. 2014).

3 Application

3.1 Anchovy survey data series

The data consist of anchovy concentration expressed in tonnes per nautical mile square

(tons/nm^2). The sampling design is the same in all years and is made of transects

perpendicular to the coast across the French continental shelf of the bay of Biscay. Transects

are regularly spaced (Fig. 2) with an inter-transect distance of twelve nautical miles (nm). The

sample values are aligned along the transects with an inter-sample distance of one nm. Fish

concentration is derived by combining echo-sounding records along the transect lines with

trawl haul catches, which allow to identify echotraces (Doray et al. 2010). Though the same transects are sailed each year their start and end are not exactly located at the same position and therefore the data locations along the transects are not positioned exactly at the same points in all years. To allow for co-kriging, the indicators of hotspots were positioned at the same points in all years. For that the data in each year were migrated on one single sampling grid for all years. The grid was made of the transects where the start was the mean start position over the years and the number of samples per transect was the mean number of samples per transect. The average difference between grid and survey sample point number along transects was 2% only. The inter-sample distance along transects was one nm. In each year, the grid node was attributed the nearest data sample.

The data sets in the different years vary over one order of magnitude in the global mean and show important differences in their aggregation pattern (Table 1, Fig. 3). The larger the global mean, the greater the maximum value, the smaller the slope of the curve Q(z) (smaller contribution of small values to the global mean).

3.2 Definition of hotspots in each year

The spatial organization was analyzed with the following suite of eight cutoffs (ranked from 1 to 8), which encompassed the range of values across years: 0.01, 10, 50, 100, 200, 400, 600, 800. The procedure defined in Sect. 2.1 was applied in each year (Table 2). In most years, the hotspot indicator corresponded to the last structured indicator. Figure 4 illustrates the variograms in this situation for year 2002. Table 2 is based on the analysis of such variograms, for all years. In 2008 and 2012 where abundance was high, the cutoff immediately succeeding the hotspot cutoff was also structured but higher values were positioned without border effect in the hotspot. This demonstrates the importance of using the variogram ratio to define hotspots. Table 3 shows the relation of hotspot's cutoff and area to

the global mean. The hotspot cutoff increased with the global mean. The hotspots varied in area occupied from one to nine percent of the study area as computed from the mean of their indicator. The occupied area of hotspots was larger for low hotspot cutoffs. Complete destructuration was observed within the hotspots (pure nugget effect of the residuals to higher cutoffs), with an exception for 2012, the richest year (residuals were structured from rank 4 defining the hotspots to rank 7, and were pure nugget effect only beyond this).

3.3 Co-kriging hotspots across years

The single variograms of the indicators of hotspots were computed in each year. No anisotropy was identified. Based on the behavior of single variograms, the coregionalization model was fitted with three predefined structural components: a nugget effect and two spherical models. The fit was performed using the model.auto() function (Dessasis and Renard 2013) implemented in the RGeostats package (Renard et al. 2014), which estimates both the sills and ranges of the structures (ranges were 13 nm and 52 nm). Table 4 shows the years where hotspots are co-regionalized (sills of cross-variograms greater than 0.002 and 0.001 in absolute value for the two structured components respectively). The hotspot areas showed consistency and variability across years. Some years (in particular 2001, 2006, 2009, 2010, 2011) were very different from the others, being correlated to a few years only. In contrast a group of years (2002, 2003, 2004, 2007, 2012, 2013) correlated with many years. Co-kriging the indicators resulted in probability maps of hotspots in each year (Figs. 5 to7). The same neighbourhood parameters were used in each year (maximum search distance of twice the inter-transect distance, minimum and maximum of samples of 5 and 10). The mean of these maps provided the fourteen year average (2000 to 2013) occurrence probability of a hotspot (Fig. 7). The location and extent of hotspots varied across years as shown by the differences in annual maps and as a result the maximum probability on the mean map was

0.24 only. Yet, there were three main areas of hotspots: off Gironde estuary, at the coast south of Gironde and on mid-shelf off Landes.

4 Conclusions

The paper shows how geostatistical non-linear indicator tools can be used to define hotspots in fisheries ecology as well as how interannual variability can be handled with multivariate geostatistics. A single cutoff was not satisfactory to define hotspots because the global mean varied across years making the concept of high value relative. Variograms and cross-variograms of indicators were used to estimate transition probabilities, which allowed to define hotspots in relative terms as the areas within which higher values occurred unpredictably. This definition is close in concept to the topcut model of Rivoirard et al. (2013). The geostatistical definition of hotspots and their mapping over time using co-kriging allowed to show convincing results on the anchovy in the Bay of Biscay. The procedure is generic and thus applicable to other fish survey data and fish species. But variability of the global mean across years and its consequence on hotspots is expected to vary between species because of specific behaviors and aggregation patterns. Here on anchovy, the global mean of the year had important consequences on the spatial organization of fish concentrations. When the global abundance was greater/smaller destructuration occurred less/more rapidly and hotspots were warmer/colder with higher/lower threshold defining the hotspots. Hotspots varied in location and extent over the years (Figs. 5 to 7, Tables 3). Yet consistent areas over time for hotspots were identified, which is useful for the conservation of habitats.

The methodology used here to define hotspots was compared to simple thresholding in each year. Bartolino et al. (2011) suggested to select a threshold based on the curvature of the cumulative distribution function (cdf) of $z(x)/\max(z(x))$. The threshold is defined as the largest value of $z(x)/\max(z(x))$ where the tangent to the cdf equals unity. The cdf shows a

curvature because low concentration values occupy large areas (high increase of the cdf) while high concentration values occupy small areas (low increase of the cdf). Thus the threshold is defined at the switch point. The treshold obtained by this procedure was systematically lower than that with our non-linear geostatistical procedure (Table 5), resulting in larger and colder hotspots. In contrast to simple thresholding, the non-linear approach used here is bivariate. It considers the behaviour of higher values within the areas defined by a lower cutoff and thus results in selecting hotspots with a larger cutoff and smaller areas. In spring during the survey period, fishing is usually located on transects 1 to 4 starting from the south (Figs. 5 to 7), which corresponded to the most sourhern hotspots identified. High abundance is an important driver of the fishing process but it is not the only one. Fish length, schooling at small scale, distance to harbour, selling price depending on harbour, fleet and social behaviour are other factors that make fishermen select some hotspots rather than others. The identification of hotspots is of interest in a conservation approach to spatially manage fisheries but may also help inform on the drivers of fishing.

References

Bartolino V, Maiorano L, Colloca F (2011) A frequency distribution approach to hotspot identification. Population Ecology 53:351–359.

Chilès JP, Delfiner P (2012) Geostatistics: modelling spatial uncertainty, 2nd edn. Wiley, New York, pp 299-385.

Desassis N, Renard D (2013) Automatic variogram modeling by iterative least squares: univariate and multivariate cases. Mathematical Geosciences 45:453-470.

Doray M, Masse J, Petitgas P (2010) Pelagic fish stock assessment by acoustic methods at Ifremer. archimer.ifremer.fr/doc/00003/11446/

Lehuta S, Mahevas S, Petitgas P, Pelletier D (2010) Combining sensitivity and uncertainty analysis to evaluate the impact of management measures with ISIS–Fish: marine protected areas for the Bay of Biscay anchovy (Engraulis encrasicolus) fishery. ICES Journal of Marine Science 67:1063–1075.

Matheron G (1982) La déstructuration des hautes teneurs et le krigeage des indicatrices. Note N-761, Centre de Géostatistique, Fontainebleau, France. cg.ensmp.fr/bibliotheque/public/

Myers N, Mittermeier R, Mittermeier C, da Fonseca G, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–858.

Nelson T, Boots B (2008) Detecting spatial hot spots in landscape ecology. Ecography 31:556–566.

Renard D, Bez N, Desassis N, Beucher H, Ors F, Laporte F (2014) RGeostats: The Geostatistical package [version 9.1.10]. MINES ParisTech. Free download from: http://cg.ensmp.fr/rgeostats

Rivoirard J (1994) Introduction to disjunctive kriging and non-linear geostatistics. Clarendon Press, Oxford.

Rivoirard J, Demange C, Freulon X, Lécureuil A, Bellot N (2013) A Top-Cut Model for Deposits with Heavy-Tailed Grade Distribution. Mathematical Geosciences 45:967-982.

Santora J, Sydeman W, Schroeder I, Wells B, Field J (2011) Mesoscale structure and oceanographic determinants of krill hotspots in the California Current: Implications for trophic transfer and conservation. Progress in Oceanography 91:397-409.

Stuart-Smith R, Bates A, Lefchek J, Duffy J et al. (2013) Integrating abundance and functional traits reveals new global hotspots. Nature 501:539-542.

Yasuda T, Yukami R, Ohshimo S (2014) Fishing ground hotspots reveal long-term variation in chub mackerel Scomber japonicus habitat in the East China Sea. Marine Ecology Progress Series 501:239-250.

Table 1: Basic statistics of the data on the single sampling grid. Years are ranked by the mean.

Mean: simple data average, CV: coefficient of variation, p0 : percent of zero values, Max:

maximum data value

Year	Mean	CV	p0	Max
2005	0.94	4.82	0.81	51
2003	2.52	4.46	0.47	275
2009	3.11	3.64	0.62	161
2007	3.33	3.13	0.54	160
2006	3.47	4.06	0.67	161
2004	4.58	5.26	0.73	420
2013	6.85	2.50	0.50	214
2010	7.04	5.08	0.73	845
2011	8.41	4.62	0.27	756
2001	9.55	4.42	0.48	635
2008	10.51	4.16	0.61	375
2000	10.89	2.27	0.28	252
2002	11.23	3.99	0.50	701
2012	17.03	3.99	0.47	1197

Table 2. Hotspot cutoffs and indicator variograms which served to define the cutoffs. Ai: hotspot defined by cutoff of rank i. Vario i: variogram of the indicator of cutoff i. Vario (i+1): variogram of the indicator of the immediately succeeding cutoff (i+1). Vario i x (i+1) /i: variogram ratio giving the spatial transition probability from cutoff i to i+1. Vario resid (i, i+1) : variogram of the indicator residual when regressing cutoff (i+1) on i. Symbol "1" indicates that the variogram was structured while "0" indicates it was a pure nugget

	Hospot Ai	Vario i	Vario (i+1)	Vario i x (i+1) /i	Vario resid (i, i+1)
2000	A4	1	0	0	0
2001	A4	1	0	0	0
2002	A4	1	0	0	0
2003	A2	1	0	0	0
2004	A2	1	0	0	0
2005	A2	1	0	0	0
2006	A3	0	0	0	0
2007	A2	1	0	0	0
2008	A4	1	1	0	0
2009	A3	1	0	0	0
2010	A4	1	0	0	0
2011	A4	0	0	0	0
2012	A4	1	1	0	1
2013	A3	1	0	0	0

Table 3: Basic statistics on hotspots. Years are ranked by the mean concentration of the year

|	Hotspot	Area proportion	Mean concentration
2005	A2	0.029	0.94
2003	A2	0.053	2.52
2009	A3	0.010	3.11
2007	A2	0.091	3.33
2006	A3	0.020	3.47
2004	A2	0.090	4.58
2013	A3	0.034	6.85
2010	A4	0.014	7.04
2011	A4	0.012	8.41
2001	A4	0.017	9.55
2008	A4	0.026	10.51
2000	A4	0.013	10.89
2002	A4	0.028	11.23
2012	A4	0.032	17.03

Table 4: Non-zero cross-correlation between the indicator of hotspots across years

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
2000	X													
2001		X												
2002		X	X											
2003			X	X										
2004	X		X	X	X									
2005				X	X	X								
2006				X	X		X							
2007	X		X	X	X	X	X	X						
2008					X			X	X					
2009			X							X				
2010			X		X			X			X			
2011												X		
2012		X	X	X	X	X						X	X	
2013	X		X	X	X	X		X			X	X	X	X

Table 5. Comparison of hotspot cutoffs resulting from a simple thresholding method with that obtained by this paper's non-linear geostatistical method. Years are ranked by the mean concentration of the year

|	Simple thresholding	Non linear geostatistics
2005	2.7	10
2003	18	10
2009	15	50
2007	14	10
2006	10	50
2004	27	10
2013	26	50
2010	61	100
2011	55	100
2001	47	100
2008	29	100
2000	40	100
2002	62	100
2012	90	100
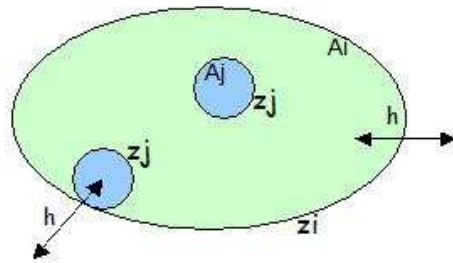
**Fig.1** Schematics showing the absence of edge effects within Ai
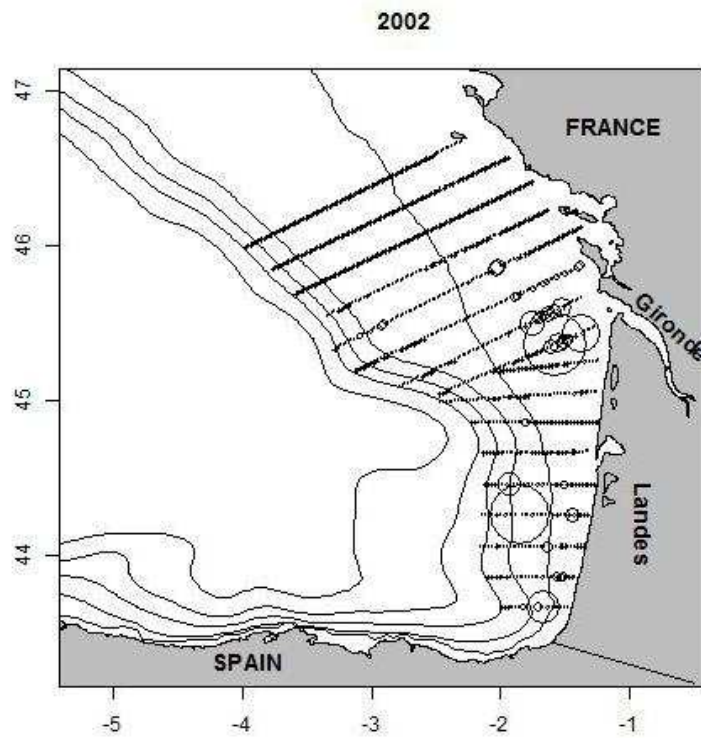
Fig.2 Design of the PELGAS surveys showing transects and sample locations. Year 2002 examplifies how high fish concentration values are located relatively to lower ones
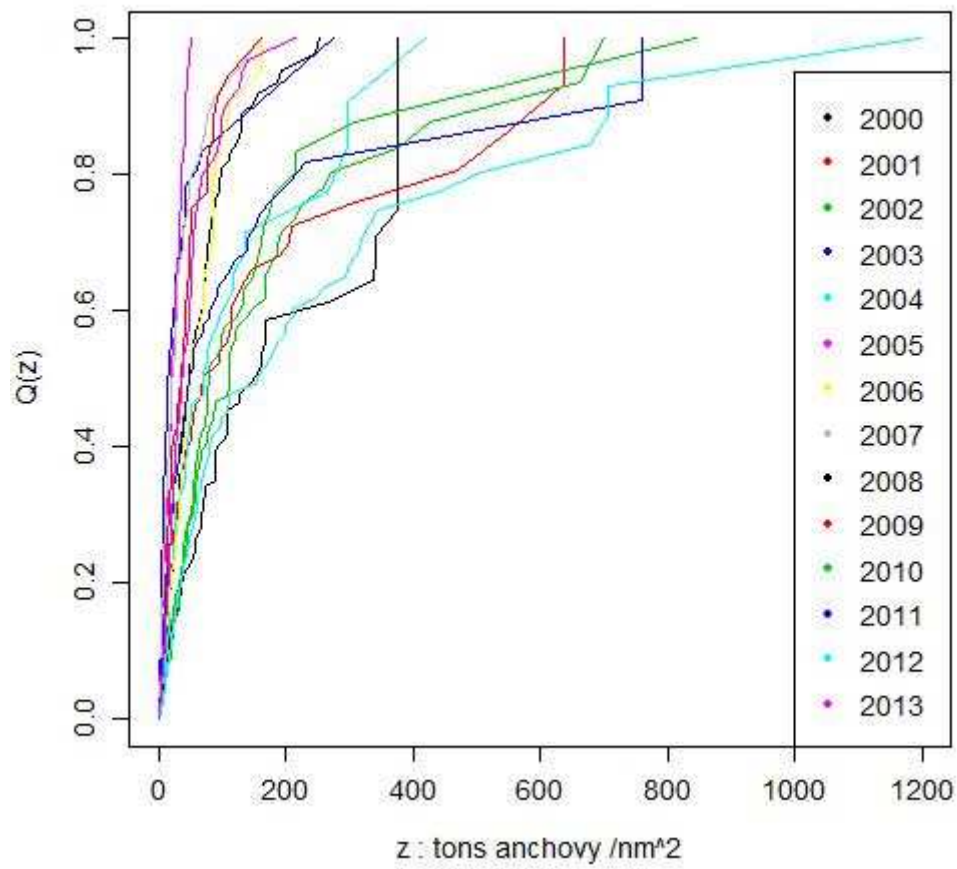
Fig.3 Selectivity curves Q(z) for years 2000-2013, where Q(z) is the summed percent biomass

as a function of fish concentration values z
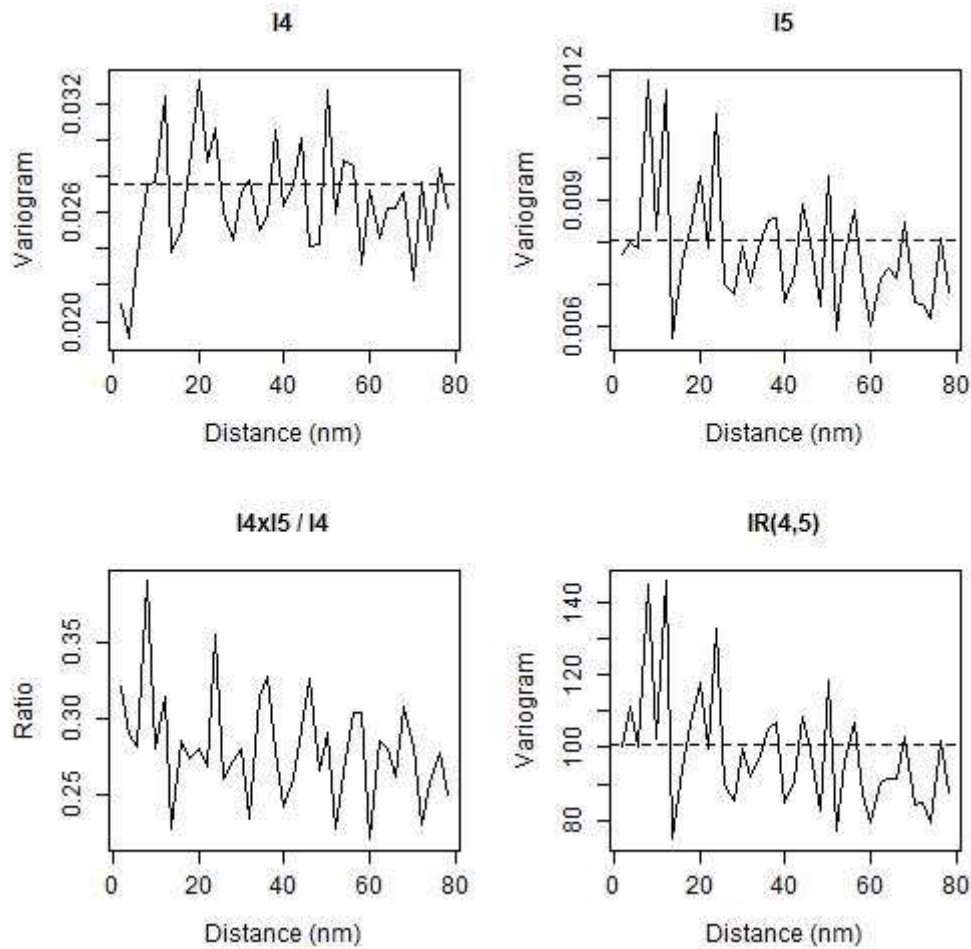
Fig.4 Variograms in year 2002 allowing to define the hotspot cutoff. The hotspot geometric set retained is A4. I4: variogram of the indicator of cutoff ranked 4 (100 tons per nm^2). I5: variogram of the indicator of cutoff ranked 5 (200 tons per nm^2). I4x5/I4: variogram ratio representing the spatial transition probability from 4 to 5. IR(4,5): variogram of the residuals of the regression of 5 on 4
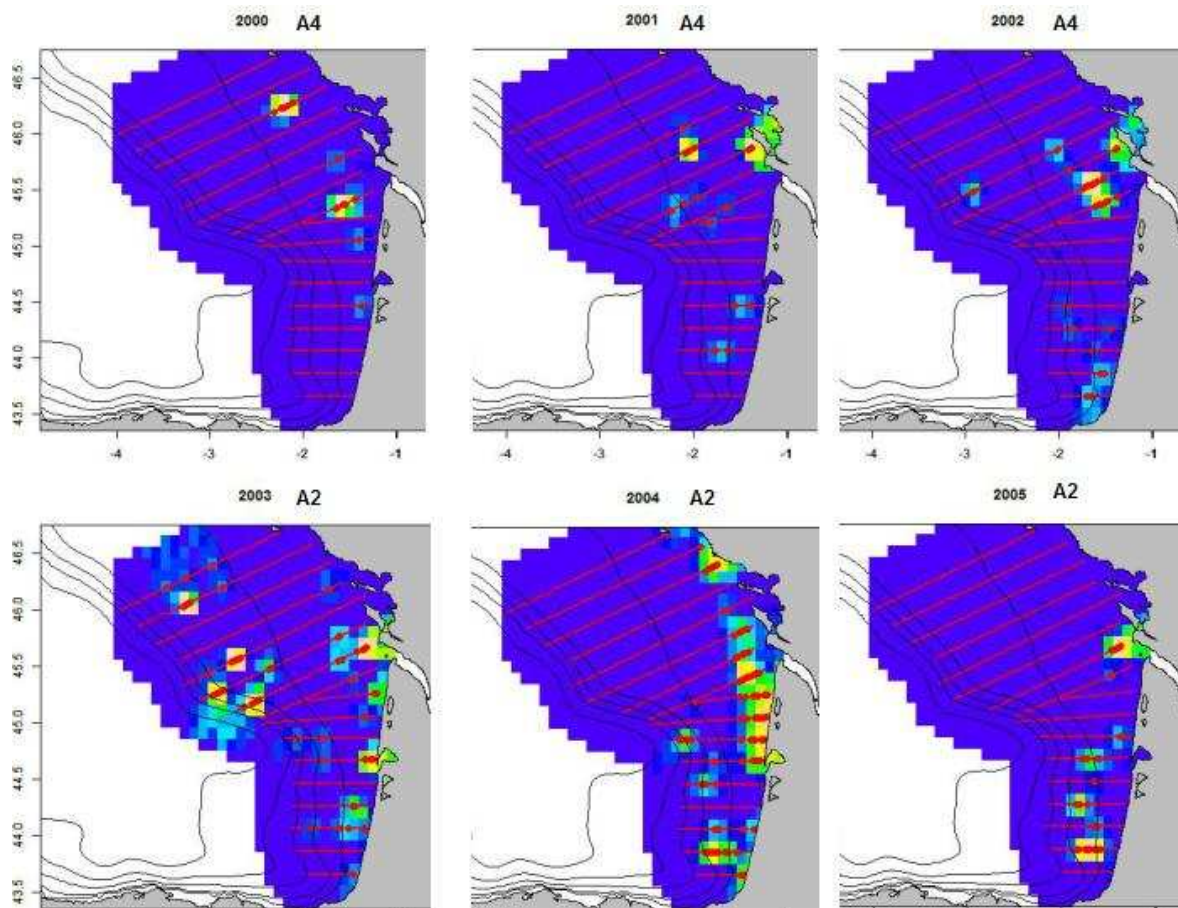
Fig.5 Probability maps of hotspots in years 2000 to 2005. The arithmetic colour scale is the same in Figs. 5 to 7 (blue: zero, red: maximum at 0.66). The along transect data are superposed as red bubbles. Ai (i in {1,…,4}) denotes the hotspot geometric set (Table 3)
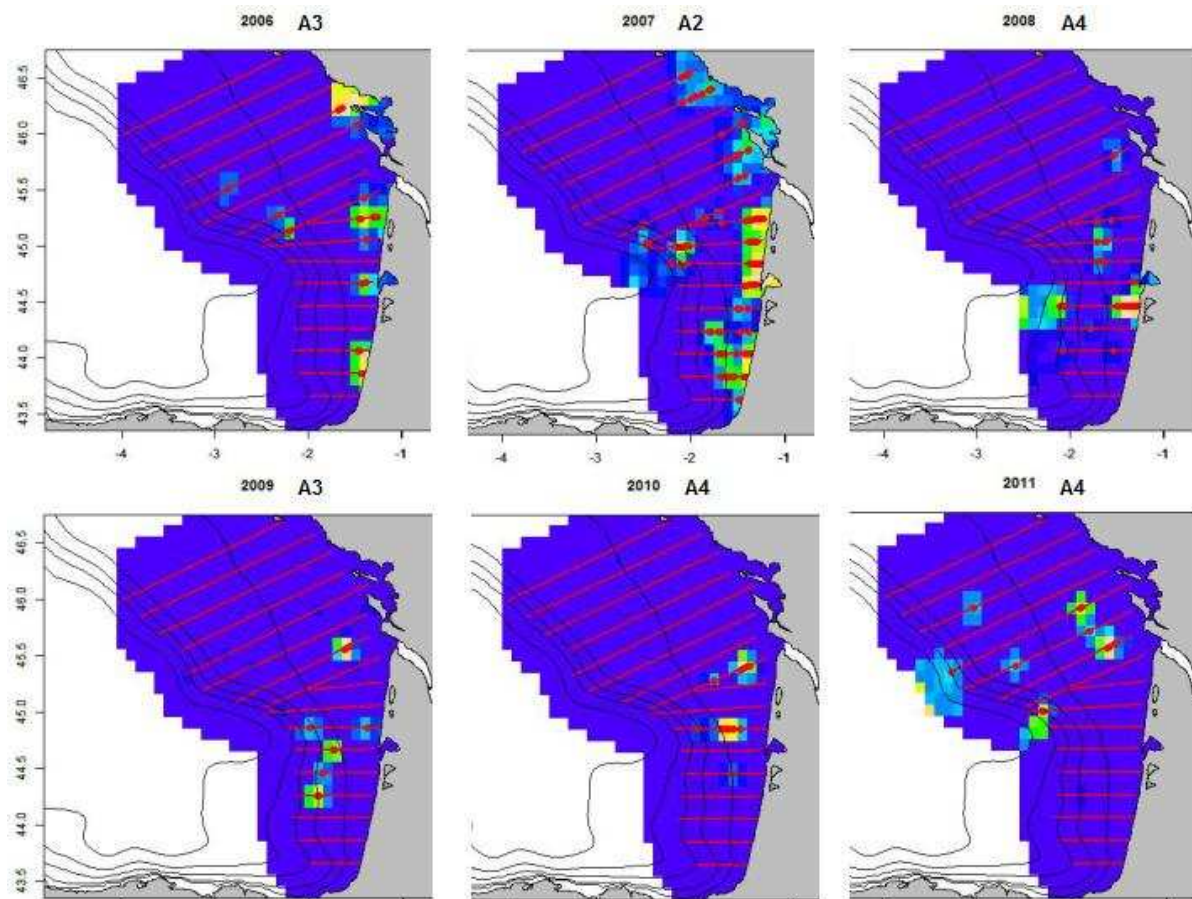
Fig.6 Probability maps of hotspots in years 2006 to 2011. The arithmetic colour scale is the same in Figs. 5 to 7 (blue: zero, red: maximum at 0.66). The along transect data are superposed as red bubbles. Ai (i in {1,…,4}) denotes the hotspot geometric set (Table 3)
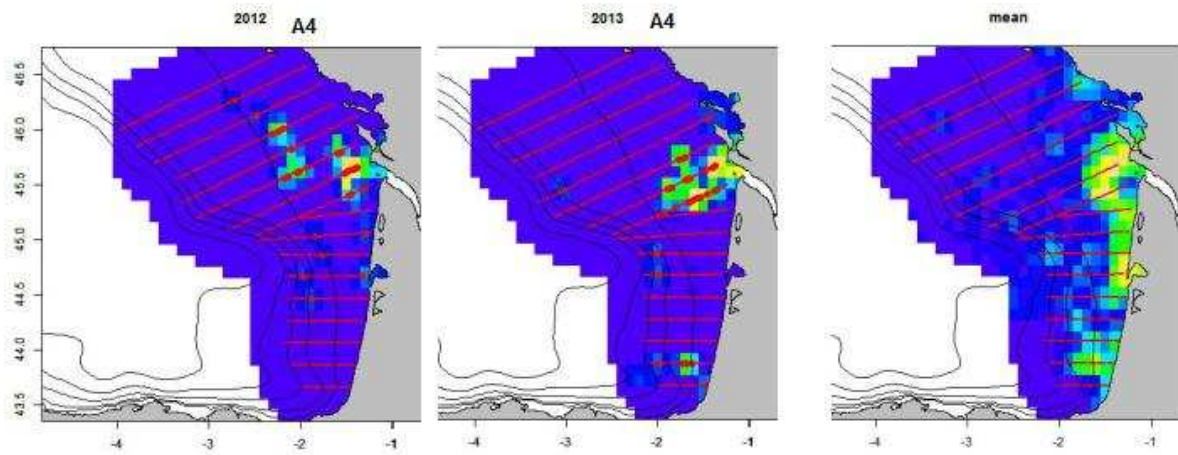
Fig.7 Probability maps of hotspots in years 2012 to 2013 and mean map for the entire series 2000 to 2013. The arithmetic colour scale is the same in Figs. 5 to 7 (blue: zero, red: maximum at 0.66). The along transect data are superposed as red bubbles. Ai (i in {1,…,4}) denotes the hotspot geometric set (Table 3)