Mathematical Geosciences
February 2018, Volume 50, Issue 2, Pages 187-208
http://dx.doi.org/10.1007/s11004-018-9725-2
http://archimer.ifremer.fr/doc/00429/54021/

Indicator-Based Geostatistical Models For Mapping Fish Survey Data

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] Centre de Geosciences, MINES ParisTech, PSL Research University, 35 rue Saint Honoré, 77305 Fontainebleau, France

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

Abstract :

Marine research survey data on fish stocks often show a small proportion of very high-density values, as for many environmental data. This makes the estimation of second-order statistics, such as the variance and the variogram, non-robust. The high fish density values are generated by fish aggregative behaviour, which may vary greatly at small scale in time and space. The high values are thus imprecisely known, both in their spatial occurrence and order of magnitude. To map such data, three indicator-based geostatistical methods were considered, the top-cut model, min–max autocorrelation factors (MAF) of indicators, and multiple indicator kriging. In the top-cut and MAF approaches, the variable is decomposed into components and the most continuous ones (those corresponding to the low and medium values) are used to guide the mapping. The methods are proposed as alternatives to ordinary kriging when the variogram is difficult to estimate. The methods are detailed and applied on a spatial data set of anchovy densities derived from a typical fish stock acoustic survey performed in the Bay of Biscay, which show a few high-density values distributed in small spatial patches and also as solitary events. The model performances are analyzed by cross-validating the data and comparing the kriged maps. Results are compared to ordinary kriging as a base case. The top-cut model had the best cross-validation performance. The indicator-based models allowed mapping high-value areas with small spatial extent, in contrast to ordinary kriging. Practical guidelines for implementing the indicator-based methods are provided.

Keywords : Top-cut, MAF, Indicators, Co-kriging, Skewed distribution, Anchovy, Bay of Biscay, Fisheries survey data, Aggregation

1 INTRODUCTION

Spatial data of natural resources or pollutants often show a small percentage of very high concentration values. This makes the inference of second-order statistics non-robust and in particular the variogram (Rivoirard et al. 2013). In fisheries survey data, high-density values often result from the fish aggregative and schooling behaviour (Fréon and Misund 1999). The aggregation dynamics occurs at a small temporal scale that the spatial survey design cannot resolve as it is designed to survey large marine areas without stopping. Thus the large data values seem to occur at random in space. When the distribution of the data is heavily skewed a usual solution is to transform the data into a better behaved, dome-shape distribution. The log transform is often considered. Yet lognormal kriging is known to be non-robust (Matheron 1974; Rivoirard 1990). Thus Guiblin et al. (1995) developed a back-transform equation to infer the variogram of the original data from that of the logged data and proceed with kriging the original data. Another solution could be transforming the data into a Gaussian distribution by normal score transforms (for treating the zero values, see Woillez et al. 2016), then performing conditional simulations and taking their average. But this remains complicated and makes a strong diffusion assumption in the spatial distribution when passing from low to high values (Matheron 1988). Another approach is to consider indicators and discretize the data. Bez and Braham (2014) discretized their data and co-kriged the corresponding indicators to map the occurrence probability of the different classes. Building on such an indicator approach, the objective of this study is to estimate the target variable using its indicators. For that, models with defined structural assumptions are considered as alternatives to ordinary kriging when the variogram is difficult to estimate. In the top-cut model (Rivoirard et al. 2013), values above a top-cut threshold have a behaviour which does not depend on lower values. This allows separating values above and below the top-cut. The

data are then mapped using co-kriging. Discretizing the target variable with multiple indicators over the full range of values was also investigated. Indicator kriging was applied, which is simple to implement but at the price of severe assumptions on the spatial structure. Another discrete model was considered, where the structure of the indicators is fully characterized and modeled. The method of min-max autocorrelation factors (MAF: Switzer and Green 1984) was applied to indicators. When the MAFs of indicators are spatially uncorrelated at all lags, they represent an empirical isofactorial model of the variable under study, which can then be mapped by kriging the MAFs separately. The objective of the paper is to compare three indicator-based models (indicator kriging, MAF, and top-cut models) together and with ordinary kriging. The models are applied to marine acoustic survey data acquired to assess the anchovy stock in the Bay of Biscay.

The data show a high proportion of zero values and a small percentage of high values. The methods are proposed as simple solutions to mapping such data and the paper gives guidelines for their implementation and discusses hypotheses that are made on the spatial distribution when using them. All calculations were performed with the package RGeostats (Renard et al. 2014) in the R environment (R Core Team 2016).

2 METHODS

2.1 Multiple indicator kriging

The target variable $Z(x)$ is discretized into p disjoint successive classes and is approximated by the discrete variable $\hat{Z}(x)$ defined by

$$\hat{Z}(x) = \sum_{k=0}^{p-1} \bar{z}_k (1_{Z(x)\geq z_k} - 1_{Z(x)\geq z_{k+1}}) \quad + \bar{z}_p 1_{Z(x)\geq z_p} \ ,$$

where $z_0=0$ and $(z_0<z_1<z_2<\ldots<z_p)$ is a set of cut-offs defining $(p+1)$ classes, $1_{Z(x)\geq z_k}$ the indicator of being above cut-off z_k and $\bar{z}_k$ the mean of $Z(x)$ in class k.

In general, it is not equivalent to krige separately the indicators of disjoint classes or the indicators of cumulated classes above cut-offs. However, when making the rather severe assumption that all indicators have the same spatial structure, those are equivalent and result in the simplest form of multiple indicator kriging. Then kriging the indicators leads to kriging the discrete variable $\hat{Z}(x)$. Thus, the variogram of the discrete variable $\hat{Z}(x)$ was computed and $\hat{Z}(x)$ was mapped by ordinary kriging. The approach approximates the continuous variable Z(x) as the intra-class variability is ignored.

2.2 MAFs of indicators as an isofactorial model

This approach is also a discrete approach where the discrete variable $\hat{Z}(x)$ (see definition above) is considered in place of the continuous variable Z(x). Here the multivariate spatial structure of the indicators is analyzed and modeled with min/max autocorrelation factors (MAF: Switzer and Green 1984). MAFs were developed to filter out noise in multi-channel (multivariate) spatial image data. The method is based on principal components analysis (PCA). MAFs are designed to maximize the autocorrelation at a given lag Δ and thus extract the most continuous information from multivariate data. Consider a p-variable random field $Z(x)=(Z_1(x),\ldots,Z_p(x))'$. The MAFs of $Z(x)$ are obtained as follows:

1. Transform with a first PCA the original data Z(x) into standardized principal components Y(x) with zero mean and unit variance.
2. Compute increments of these principal components for lag Δ, Y(x)-Y(x+Δ), and apply a PCA on their covariance matrix
3. Re-order the principal components of this second PCA with eigenvalues in increasing order. The (re-ordered) principal components are the MAFs $\chi(x)$.

By construction, the MAFs have zero mean and are uncorrelated at lag 0 and also at lag Δ. The first MAF has the highest autocorrelation (smallest variogram value at lag Δ), the second

MAF the second highest and so one. In fisheries ecology, MAFs have been applied on multiple time series (Fujiwara 2008; Woillez et al. 2009) to extract the most continuous series and regroup series with common behaviors. Following Rivoirard et al. (2014), MAFs were used here in the two-dimensional geographical space to model and map the continuous variable *Z(x)* with a discrete approach. The target variable *Z(x)* was also discretized into *p* disjoint successive classes and the corresponding discrete variable $\hat{Z}(x)$ written as above

(Sect. 2.1) $\hat{Z}(x) = \sum_{k=0}^{p-1} \bar{z}_k (1_{Z(x) \geq z_k} - 1_{Z(x) \geq z_{k+1}}) + \bar{z}_p 1_{Z(x) \geq z_p}$.

The indicators of the *(p+1)* classes represent a multivariate data set, from which *p* MAFs were extracted. When the MAFs of indicators are spatially uncorrelated at all lags (not just at lags 0 and Δ), they form an empirical isofactorial decomposition of the discrete variable (Rivoirard et al. 2014) $\hat{Z}(x) = m + \sum_{i=1}^{p} c_i \chi_i(x)$, where $m=E[Z(x)]$ is the mean of *Z(x)*, $\chi_i(x)$ are the zero mean MAFs of indicators and c_i are coefficients. The coefficients c_i are derived from the relationships $E[\hat{Z}(x)\chi_i(x)] = c_i E[\chi_i(x)^2]$ as the MAFs $\chi_i(x)$ and $\chi_j(x)$ $(i \neq j)$ are spatially uncorrelated.

Kriging the discrete variable $\hat{Z}(x)$ is then obtained by kriging the MAFs of indicators separately and by linearly combining the kriged estimates. Higher order MAFs showing pure nugget effects may be kriged or not (see below). Spatial uncorrelation between MAFs is checked by computing cross-variograms between them.

2.3 Top-cut model

Here, the discretization approach applies to high values only. In the top-cut model (Rivoirard et al. 2013) the target variable *Z(x)* is truncated at a defined sufficiently high threshold (the top-cut cut-off z_e) and values above it are treated separately. Variable *Z(x)* is split into three

components, $Z_1(x)$ the truncated variable $Z(x)^{\wedge}z_e$, $Z_2(x)$ the weighted indicator of the threshold $(m(z_e)-z_e)\ 1\{Z(x) \geq z_e\}$ and $Z_3(x)$ the residual $Rz_e(x)$, which add

$Z(x) = Z(x)^{\wedge}ze + (m(z_e)-z_e)1\{Z(x) \geq z_e\} + Rz_e(x)$. The truncated variable is $Z(x)^{\wedge}ze = Z(x)$ if $Z(x)<ze$ and ze otherwise. The indicator of the top-cut cut-off is $1\{Z(x) \geq z_e\}=0$ if $Z(x)<ze$ and 1 otherwise. The mean of Z above the cut-off is $m(z_e)$ and therefore $(m(z_e)-z_e)1\{Z(x) \geq z_e\}$ represents the mean excess variable (mean excess value times the indicator). And the residual $Rz_e(x) = [Z(x) - m(ze)]1\{Z(x) \geq z_e\}$ represents the variability around the mean of Z when above z_e.

In this model, the truncated variable $Z_1(x)$ and the weighted indicator $Z_2(x)$ are spatially correlated. The residuals $Z_3(x)$ may show a spatial structure or not. The top-cut model considers that the residual $Z_3(x)$ is spatially uncorrelated with the truncated variable $Z_1(x)$ and the indicator $Z_2(x)$. Mapping $Z(x)$ is then obtained by co-kriging the truncated variable $Z_1(x)$ and the weighted indicator $Z_2(x)$. The zero mean residual $Z_3(x)$ may be kriged and added to the other two components or not (see below). In this latter case high values are accounted for by the mean excess value $(m(z_e)-z_e)$, a constant, which adds to the truncated variable but only in areas where the probability to exceed the threshold is high as defined by the estimation of the indicator $1\{Z(x) \geq z_e\}$. The spatial structures necessary for mapping are the simple- and cross-variograms of the truncated variable $Z_1(x)$ and the weighted indicator $Z_2(x)$, which are inferred without considering the variability in the very high values.

The major assumption in the model is the absence of border effect in the spatial organization of the residual $Z_3(x)$ within the geometrical set defined by the indicator $1\{Z(x) \geq z_e\}$. In other words, the residual values are located inside this set with no influence on the proximity of its borders. If there was a border effect, smaller values would be close to the borders and larger values in the center of the set, or vice versa. This can be analyzed using the ratio of the cross-variogram between the residuals and the indicator divided by the variogram of the indicator

(Rivoirard et al. 1994, 2013). As a matter of fact, this variogram ratio measures the average value of $(Z(x)- m(z_e))$ when entering the geometrical set defined by the indicator $1\{Z(x) \geq z_e\}$. A flat ratio indicates an absence of border effect. In practice, the top-cut cut-off z_e is chosen as the lowest cut-off starting from the high values for which the residual $Z_3(x)$ values do not exhibit any border effect within the geometrical set defined by the indicator.

2.4 Kriging pure nugget components

The top-cut and MAF models separate different components in the spatial distribution, some of which may be zero mean pure nugget effects. Should they be kriged? When kriging with the top-cut model, the kriged estimate of a block v_0 is

$$Z^{ktc}(v_0) = Z_1^{ck}(v_0) + Z_2^{ck}(v_0) + Z_3^{k}(v_0),$$

and its estimation variance

$$\sigma_{ktc}^2(v_0) = \sigma_{ck,1+2}^2(v_0) + \sigma_{k,3}^2(v_0),$$

where *ck* stands for co-kriging, *k* for kriging and *ktc* for kriging with the top-cut model and where the numbers in subscript indicate the components of the model.

Component Z_3 (residuals) has zero mean by construction. If it corresponds to a pure nugget effect, one may wonder whether or not the component should be kriged. If kriging of Z_3 is omitted, one considers that in the neighborhood of the block, the mean of Z_3 is zero and thus assumes stationarity in the high values above the cut-off, which results in replacing them by their constant mean $m(z_e)$. This may be a strong hypothesis but it is also a practical one when high values are very imprecisely known. In that case, values above the top-cut cut-off are modeled by component Z_2 only, that is, they are estimated by the probability of being above the cut-off (indicator) multiplied by a constant, the mean excess value. In this case, the nugget component does not contribute to the estimation variance of the block average $\sigma_{k,3}^2(v_0) = 0$.

Alternatively, kriging Z_3 will estimate local residual variations around the mean $m(z_e)$ in areas above the top-cut cut-off. One should, therefore, decide whether or not it is reasonable to represent such details because of the imprecise knowledge on the high values. When kriging the residuals, the corresponding estimation variance is $\sigma_{k,3}^{\ 2}(v_0) = \mathrm{var}(Z_3)/n(v_0)$, where $n(v_0)$ is the number of samples in the neighborhood. It adds to the estimation variance of the other components.

Similarly, for a fitted MAF isofactorial model with p components, the last q $(q<p)$ may be pure nugget effects. When kriging block v_0, the estimate of the block average is

$$\hat{Z}^{kmf}(v_0) = m + \sum_{i=1}^{p-q} c_i \chi_i^{\ k}(v_0) + \sum_{j=p-q+1}^{p} c_j \chi_j^{\ k}(v_0),$$

and its estimation variance

$$\sigma_{kmf}^{\ 2}(v_0) = \sum_{i=1}^{p-q} c_i^{\ 2} \sigma_{ki}^{\ 2}(v_0) + \sum_{j=p-q+1}^{p} c_j^{\ 2} \sigma_{kj}^{\ 2}(v_0),$$

where *kmf* stands for kriging with the MAF model and *k* for kriging.

The MAFs have zero mean. Therefore the q last MAFs showing nugget effects may not be kriged. Their kriging may add noise or provide local details in the spatial distribution, depending on our capacity to interpret this signal. As discussed above, the nugget components add no variance to the estimation of the block average when they are not kriged, whereas they contribute by $\sigma_{kj}^{\ 2}(v_0) = \mathrm{var}(MAF_j)/n(v_0)$ when they are. Here kriging was performed with and without the pure nugget components for comparison.

The top-cut, the MAFs of indicators and the multiple indicator models were fitted on the anchovy data (see below) and were also compared to ordinary kriging, which was used as a base case for model comparison. Therefore, the variogram of Z(x) was also calculated.

Kriging with the four models was performed on the same grid and with the same moving neighborhood.

2.5 Comparison of approaches

To compare model performances, the prediction errors associated with each model were compared. Prediction errors were estimated by cross-validation, that is, considering each sample as temporarily unknown and comparing the sample value to the one estimated by kriging. Prediction errors were characterized using global statistics such as the overall bias and the mean squared error. But these statistics may be non-robust because of some high or low values difficult to re-estimate. Thus following Rivoirard et al. (2013), the regression of the sampled values on the estimated ones was used as well, to characterize whether values estimated in a given range were effectively observed in that range (this refers to non-conditional bias).

To compare the patterns in the spatial distributions estimated with the different models, the kriged maps were compared. This was performed using visual comparisons, selectivity curves and scatter plots. Selectivity curves (Matheron 1981) allowed comparing the dispersion in the estimated values for the different models. The selectivity curve is defined by $Q(z)=E[Z\ 1_{Z\geq z}]$, where Z stands for the kriged estimate. Scatter plots were used to compare the estimated values among the different models.

3 APPLICATION

3.1 Anchovy survey data

The data (Fig. 1) consisted of anchovy density values expressed in tons per nautical mile square (tons nm^{-2}). The sampling design was made of line transects perpendicular to the coast across the French continental shelf of the Bay of Biscay. Transects were regularly spaced with

an inter-transect distance of twelve nautical miles (12 nm). The inter-sample distance along transect was one nm. Fish density was derived from the combination of echo-sounding records and trawl-haul catches undertaken to identify the echo-traces (Doray et al. 2010). The 2002 data showed a marked zero effect (50% of the data) and a few high values with a disproportionate contribution to the data arithmetic mean (3% of the data were above 100 tons mn^{-2} and represented 45% of the mean, Table 1). Some of the high values were spatially aggregated while others appeared isolated in areas of low values. The proportion of very low and low values was 40% and that of medium values 3% (Table 1). The basic statistics were: 985 samples, with an arithmetic mean of 11.23, a coefficient of variation of 3.99, a maximum value of 701, and 50 percent of zero values.

3.2 Structural analysis

3.2.1 Variogram

The experimental variogram of the data was computed in two directions (not shown) along and across transects. No anisotropy was identified. The omnidirectional variogram with a lag of 2 nm (+/- 1 nm) showed a rapid increase and large fluctuations around a sill (Fig. 2). It was modeled with in a high nugget effect (65% of total sill) and a spherical model with a short correlation range (9 nm). Yet, a larger correlation structure is visible (Fig.1) for medium values.

3.2.2 Discretizing and thresholding the data

The data were discretized into five classes on a scale close to the decimal log scale (Table 1). The threshold defining the last class was chosen as equal to the top-cut cut-off in the top-cut fitted model. After various trials, the top-cut cut-off of a 100 tons nm^{-2} was retained as it allowed a consistent fit of both top-cut and MAF models. Adding higher classes in the MAF analysis to the current five classes resulted in estimating unnecessary higher order MAFs,

which were not spatially structured. Hence the current choice of classes, which will be discussed further. The mean of the discretized data $\hat{Z}(x)$ was equal to that of Z(x) (11.23 tons nm^{-2}) whereas the coefficient of variation of $\hat{Z}(x)$ was reduced in comparison to that of *Z(x)* even if it still remained large (3.26 instead of 3.99).

3.2.3 Variogram of the discrete variable

The experimental variogram of the discrete variable $\hat{Z}(x)$ was computed and modeled in the same way as the continuous variable Z(x) (Fig. 3). The correlation range of $\hat{Z}(x)$ was larger than that of Z(x) and the nugget effect remained high. The correlation range was similar to that of the second MAF or that of the truncated variable Z(x)^ze. The shorter correlation range of Z(x) was thus due to the presence of high values, whose effect was reduced when discretizing the data.

3.2.4 MAFs of indicators as an isofactorial model

MAFs were computed on the multivariate data set made of the indicators of the five classes used to discretize the anchovy data (Table 1). With five classes, four MAFs were estimated. They were computed using different omnidirectional lags Δ = 1, 2, 5, 12 nm. To characterize the sensitivity of MAFs with respect to the lag Δ, mean MAF values were reported as a function of the class index (Fig. 4). The mean MAF value for a given data class was calculated on the samples belonging to that class. MAFs 1 and 2 were not very sensitive to the lag value Δ, indicating that the most continuous components in the spatial distribution were robustly defined. Hence a lag Δ = 2 nm was used throughout this study. The curves of mean MAFs as a function of class indices allowed interpreting the MAFs (Fig. 5). MAFs with a large absolute mean value for a given class index represented that class in particular. A

gradual change in the mean MAF value with the class index indicated smooth spatial transitions between classes. MAF 1 revealed a gradual spatial transition from class 1 to class 3. MAF 2 characterized class 2 and class 5 in particular. MAF 3 accounted for the spatial difference between class 5 and classes 3 and 4 and MAF 4 between classes 3 and 4. Omnidirectional simple and cross-variograms of the MAFs of indicators were computed with a lag of 2 nm. The variograms of MAF 1, MAF 2 and MAF 3 were structured with decreasing autocorrelation range and that of MAF 4 showed a pure nugget effect (Fig. 6). Cross-variograms between MAFs were unstructured (Fig. 7), meaning that the MAFs constituted an isofactorial model of the discretized anchovy data.

3.2.5 Top-cut model

Omnidirectional simple and cross-variograms were computed with a lag of 2 nm. The variogram of the indicator $I_e(x)=1\{Z(x) \geq z_e\}$ was structured (Fig. 8) with a correlation range close to that of the variogram of the data, meaning that the values above the top-cut cut-off strongly influenced the overall spatial structure. The residual $Z_3(x)$ was a pure nugget effect component and showed no border effect with the indicator $I_e(x)$, as demonstrated by the variogram ratio which was flat (Fig. 8). The top-cut model assumption was thus validated: within the geometrical sets defined by the indicator, the high values above the cut-off occurred at random and without transition (no border effect) when crossing the spatial limits of the sets.

The variogram of the truncated variable $Z_1(x)$ revealed a longer range structure (20 nm) (Fig. 8), than the variogram of the data. The variogram of the mean excess variable $Z_2(x)$ is by construction proportional to that of the indicator and showed a shorter range structure, close to that of the variogram of the data. This result was consistent with the so-called de-structuration of the high values (Matheron 1982), where indicators for increasing cut-offs have decreasing

correlation ranges. A linear model of co-regionalization between $Z_1(x)$ and $Z_2(x)$ was fitted using the algorithm of Goulard and Voltz (1992). The model was composed of three basic structures: a nugget effect and two spherical models with short (8nm) and long (25nm) range structures (Fig. 9).

3.3 Choice of neighborhood

Because they were collected along transects, the data were not evenly distributed in all spatial directions and this could affect kriging results. The choice of the neighborhood was thus critical. For kriging between transects, data from different transects and quadrants were used. Different neighborhoods were tested by cross-validation using ordinary kriging. The neighborhood retained gave the best cross-validation statistics, in particular, the lowest estimation bias and a ratio between kriging standard error and estimation standard error closest to one. The retained neighborhood was defined as follows: radius of 30 nm, four samples at minimum and sixteen at maximum with a maximum of four samples by quadrant.

3.4 Cross-validation and comparison of model performances

Cross-validation results varied slightly between models (Table 2). The overall bias was small for all models (0.1 to 0.5% of the mean). The root mean squared error was comparable between models, although ordinary kriging had the largest value. The predicted kriging error was lower for all models than that observed by cross-validation (Table 2: error ratio between 1.5 and 1.9). This was due to the underestimation of high values in the vicinity of low values and vice versa. To appreciate whether values estimated in a given range were originally observed in that range, the class means of the re-estimated samples were computed and compared to that of the corresponding samples (Fig. 10). The points should be close to the diagonal line, which corresponds to the non-conditional estimation bias. The values estimated

in the lower ranges (<50 tons nm^{-2}) were originally low on average for all models. Models differed for values estimated in the highest class (>100 tons nm^{-2}). For the models based on indicators only (MAF and indicator kriging), values estimated in that class underestimated real values. Ordinary kriging showed overestimation, probably because of an overestimation of low values in the vicinity of high ones (see later). And the top-cut model estimates were close to the diagonal line, meaning that this approach estimated the real values in that highest class correctly on average. All models overestimated the real values in the intermediate range (50-100 tons nm^{-2}), although the top-cut and MAF models without their nugget components were closer to the diagonal line.

3.5 Mapping by kriging

The data were mapped by block kriging. Block kriging was performed on a grid with a mesh size of 6 nm, which encompassed the survey design and was limited by the coastline and the polygon of species potential presence offshore. Maps were derived by ordinary kriging, indicator kriging, kriging with the MAF and top-cut models (Fig. 11). A visual inspection showed that in ordinary kriging, high data values spread their influence on the estimation in their neighborhood. This effect could also be seen with the top-cut model with residuals but was not seen with the other models, which then estimated better the small spatial extent of rich values in regions of low to medium values. With these latter models, the surroundings of high isolated data values were not estimated as rich areas (as for instance on the fourth transect counted from the southern limit). Also, with all models except ordinary kriging, areas surrounding medium data values were estimated as such (as for instance on the first and second transects in the South or offshore on the fifth transect counted from the northern limit). The scatter plots between map values obtained with the different models (Fig. 12) allowed identifying more precisely similarities and dissimilarities between models. Results obtained

with indicator kriging, MAF, and top-cut models without residuals were correlated, while the top-cut model with residuals was better correlated to ordinary kriging. It is noteworthy that some intermediate values estimated around 50 (tons nm^{-2}) with ordinary kriging were estimated within a larger range (20-80) with indicator kriging, MAF and top-cut without residuals, probably due to the larger spatial continuity in these models for low and intermediate values (range 10-100 tons nm^{-2}: class indices 3 and 4). Furthermore, the selectivity curve for ordinary kriging (Fig. 13) contrasted with that of other models, being more selective in the range 40-80 (tons nm^{-2}). This showed again how intermediate values played a major role in determining the structure and the spatial pattern in the kriged maps, when using indicator kriging, MAF and top-cut models without residuals.

Basic statistics of the kriged maps are presented in Table 3. The means were similar for ordinary kriging and the top-cut model with residuals on the one hand, and for the other models on the other hand, the former being smaller than the latter. The highest values estimated by ordinary kriging and the top-cut model with residuals were the highest among models. The highest values estimated by MAF and indicator kriging were the lowest among models, probably because the spatial structure in these models was more continuous. For the MAF and top-cut models, the maximum (respectively minimum) estimated values were greater (respectively lower) when nugget components were kriged. The negative estimated values had a low impact (<1%) on the means.

The maps of standard deviation (kriging error) predicted by the different models showed similar spatial patterns: lower precision on the borders of the field as well as in between transects. This was so because kriging with the different models was performed with the same neighborhood on the same grid. Thus only mean kriging errors were compared (Table 3). Ordinary kriging had the highest mean kriging error. In the top-cut model, the residuals had a high variance, which led to an important additional error when comparing kriging with and

without residuals. Indicator kriging and kriging with the MAF model showed similar kriging errors. The difference between ordinary kriging and other models resulted from the fact that the variogram of the data had a high nugget effect, high sill, and short correlation range because of a few high values. In contrast, the spatial structures in the other models had longer ranges representing a greater proportion of total variance and were less affected by high values.

4 CONCLUSION - DISCUSSION

When data have a small proportion of high values, the experimental variogram is often erratic, poorly structured, and its modelling uncertain. The paper showed how geostatistical indicator-based models can be used to map such data. The top-cut, MAF of indicators and multiple indicator kriging models were alternative efficient approaches to ordinary kriging. In these approaches, the residual variability in the class of largest data was erased, and the spatial structure in the low and medium values served to guide mapping. The spatial structures in these models had a longer range and represented a greater proportion of total variance than in the variogram used for ordinary kriging. The practical implementation of the models was based on discretizing the data in classes (introducing thresholds) and a thorough structural analysis. Model performances were analyzed by cross-validation the data and comparing the kriged maps. The top-cut without residual model had the best cross-validation performance (lowest mean squared error and no conditional bias). In the cross-validation exercise, the highest data values were underestimated when using MAF and indicator kriging models. All models overestimated real values in the medium range of values whereas estimated low values were effectively low. Comparing spatial patterns in the kriged maps showed that the top-cut without residual, MAF and indicator kriging approaches were all able to map high-value areas with small spatial extent even when surrounded by low values. In contrast, the

map obtained by ordinary kriging showed a spatial spread in the influence of high values on low ones and vice versa.

4.1 Guide lines for model choice

The top-cut model, when applicable, appears to be a convenient way to handle high values, when they are not frequent but make statistics non-robust. Very often the residual is a pure nugget effect (at least locally, since the variograms are generally not computed for large distances). Two options are then available for the estimation of the initial variable. Either kriging the residuals assuming local stationarity in the neighborhoods: then the estimates depend on the high data values present in the neighborhoods; or using the zero mean residual value: this makes the estimate more robust as it does not depend on individual high data values. But this assumes strict stationarity of the residuals, meaning that the values higher than the top-cut cut-off have the same mean everywhere in space.

MAFs of indicators extract the most spatially structured components, which are most helpful for mapping. However, this approach first requires a discretization of the variable. In doing so, a part of the variability disappears. In particular, the data values in the highest class are set to their mean, which is considered to be valid everywhere in space. On the other hand, the last MAF is likely to be a pure nugget effect. Then, it can either be kriged or not. In this latter case, its zero mean is used instead, assuming it is valid everywhere.

In the present application on anchovy survey data, high values corresponded to dense schools, with no understanding of their dynamics. Therefore, the accurate location of the high values as well as their maximum value were imprecise. For these reasons, it seems appropriate here not to krige the pure nugget components neither in the top-cut nor in the MAF models. In other situations where the process generating the high values would be better known, kriging the pure nugget components could be worthwhile.

In our example, multiple indicator kriging and MAF kriging performed similarly, probably because the variogram of the discrete variable was close to that of the first two MAFs. It is difficult to judge whether this result is generic or case dependent. The MAF approach could thus be recommended for analyzing precisely the spatial structure in the discretized data. Table 4 summarizes the different steps when analyzing data with the top-cut and MAF models.

4.2 Thresholding the data

Rivoirard et al. (2014) suggested defining those discretization classes that lead to the best approximation of the data selectivity curve. In doing so, one is tempted to add extra classes for the high values, with low probability. This results in diminishing the reliability of the first MAFs and increasing the number of high order MAFs with pure nugget effect models. Therefore, here, only five data classes were considered and high values were regrouped in one class. As a result, only the highest order MAF showed a pure nugget effect. The drawback was a slight decrease in variance of the discretized variable in comparison to the data and a lower maximum estimated value. The advantage was the estimation of robust MAFs (robust to the discretization and to lag Δ). They revealed long-range spatial structures in the data, which were used for mapping.

4.3 Behaviour of the first MAF of indicators as a function of the class index

The distribution of MAF 1 as a function of the class index used when discretizing target variable $Z(x)$ may be used as an empirical diagnosis for choosing between two classes of models: the diffusion or the sharp transition (jump) types (Matheron 1988; Rivoirard 1994). A monotonic curve is indicative of gradual spatial transitions from low to high values, which would be adequately modeled by a diffusion model (Rivoirard et al. 2014). In our case (Fig.

5) the variation of MAF 1 with increasing class index showed a decrease from class 1 to class 3 and then a sill for higher class indices. Thus MAF 1 indicated a gradual spatial transition between data class indices 1 to 3 but such diffusion in the low values did not extend to higher class indices 4 and 5. The sharp spatial transitions to high values (class 5) were identified by the absence of border effect in the top-cut model and this behaviour was apparent in MAFs 2 and 3. The structural analyses in the MAF and top-cut model approaches both identified a mixed spatial behaviour: gradual and structured spatial transitions in the low values, on top of which high values occurred with very sharp transitions and no spatial structure. This switch in spatial behaviour occurred above 100 tons nm^{-2}. The biological implications of such density threshold for anchovy aggregative behaviour remain to be understood.

ACKNOWLEDGEMENTS

We are grateful to the crew of the research vessel Thalassa and to E. Duhamel, F. Sanchez and P. Grellier from Ifremer for preparing the biological and acoustic data. We would also like to thank the referees, whose comments allowed to improve the manuscript. D. Renard (Mines Paris Tech) helped to improve the English. The work was partly funded by the European Union H2020 project CERES. The data were collected by Ifremer within the French national observation plan, a part of the EU fisheries data collection framework.

REFERENCES

Bez N, Braham C (2014) Indicator variables for a robust estimation of an acoustic index of abundance. Canadian Journal of Fisheries and Aquatic Sciences 71: 709–718

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

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

Fréon P, Misund O (1999) Dynamics of pelagic fish distribution and behaviour: effects on fisheries and stock assessment. Blackwell Science, Oxford

Goulard M, Voltz M (1992) Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. Mathematical Geology 24:269–286

Guiblin P, Rivoirard J, Simmonds E (1995) Analyse structurale de données à distribution dissymétrique : exemple du hareng écossais. Cahiers de Géostatistique 5: 137-160. http://cg.ensmp.fr/bibliotheque/

Matheron G (1974) Effet proportionnel et lognormalité ou : le retour du serpent de mer. Note N-374, Centre de Géostatistique de Fontainebleau. http://cg.ensmp.fr/bibliotheque/

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

Matheron G (1988) Two classes of isofactorial models. In: Armstrong M (ed) Geostatistics, Kluwer Academic Publishers, pp 309-322

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Austria. http://www.R-project.org

Renard D, Bez N, Desassis N, Beucher H, Ors F, Laporte F (2014) RGeostats: The Geostatistical package [version 11.0.2]. MINES ParisTech. http://cg.ensmp.fr/rgeostats

Rivoirard J (1990) A review of Lognormal estimators for in-situ reserves. Mathematical Geology 22: 213-221

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

Rivoirard J, Freulon X, Demange C, Lecureuil A (2014) Kriging, indicators and non-linear geostatistics. The Journal of the Southern African Institute of Mining and Metallurgy 114: 1-6

Switzer P, Green A (1984) Min/Max autocorrelation factors for multivariate spatial imagery. Technical Report No. 6, Department of Statistics, Stanford University: 14p

Woillez M, Rivoirard J, Petitgas P (2009) Using min/max autocorrelation factors of survey-based indicators to follow the evolution of fish stocks in time. Aquatic Living Resources 22: 193-200

Woillez M, Walline P, Ianelli J, Dorn M, Wilson C, and Punt A (2016) Evaluating total uncertainty for biomass- and abundance-at-age estimates from eastern Bering Sea walleye pollock acoustic-trawl surveys. ICES Journal of Marine Science 73: 2208–2226

Table 1: Parameters of the discrete distribution resulting from discretizing the anchovy data

Values	Class index	Limits	Proportion	Mean
zeroes	1	[0,0.1[	0.539	0.0003
very low	2	[0.1,10[	0.296	2.55
low	3	[10,50[	0.109	23.94
medium	4	[50,100[	0.027	67.67
high	5	[100,+∞ [	0.028	211.26

Table 2: Statistics of cross-validation for the different models. *Zest* are the re-estimated values, *Zobs* the sampled values and *sigma.K* the estimation error predicted by kriging. Models are: ordinary kriging (OK), indicator kriging (IK), MAF model without (estim1) and with (estim2) the nugget component, Top-cut model without (estim1) and with (estim2) the nugget residual

	OK	IK	MAF estim1	MAF estim2	Topcut estim1	Topcut estim2
Mean estimation error *mean [Zest-Zobs]*	0.01	0.01	-0.04	-0.05	-0.04	-0.05
Root mean squared error *√ mean [[Zest-Zobs]^2]*	40.75	40.08	40.00	40.13	39.54	40.16
Mean normalized error *mean [|Zest-Zobs| / sigma.K]*	1.56	1.68	1.94	1.87	1.51	1.48

Table 3: Statistics of the maps obtained by block kriging with the different models. The data arithmetic mean is 11.23 and the standard deviation 44.86. Models are: ordinary kriging (OK), indicator kriging (IK), MAF model without (estim1) and with (estim2) the nugget component, Top-cut model without (estim1) and with (estim2) the nugget residual

	OK	IK	MAF estim1	MAF estim2	Topcut estim1	Topcut estim2
Mean	10.40	11.05	11.10	10.97	10.96	10.48
Mean kriging error	43.86	32.44	30.89	32.34	25.00	36.39
Maximum	171.77	124.97	111.48	121.32	128.45	146.39
Minimum	0	0.003	-2.87	-4.44	-2.37	-3.37

Table 4: The different steps when analyzing survey data with the top-cut and MAF models

	Top-cut model	MAF of indicators model
Target variable and its decomposition	Variable *Z(x)* *Z(x)* is decomposed into three components: truncated variable, indicator of being above the cutoff and residual	Discretized variable $\hat{Z}(x)$ $\hat{Z}(x)$ is decomposed into MAFs
Model assumptions	Values above the top-cut cut-off are spatially uncorrelated with the indicator	MAFs of indicators are spatially uncorrelated between each others MAFs of indicators are robust against changes in lag Δ used to compute them
Implementation steps	Identify a cut-off value that satisfies the previous assumption Compute simple and cross variograms of the components and check whether previous assumption is valid. If not, use MAFs	Discretize Z(x) using a sufficient number of classes Compute MAFs for different lags Δ and evaluate robustness of MAFs against the choice of Δ Compute simple and cross variograms of MAFs and check if previous assumption is valid. If not, consider indicator kriging

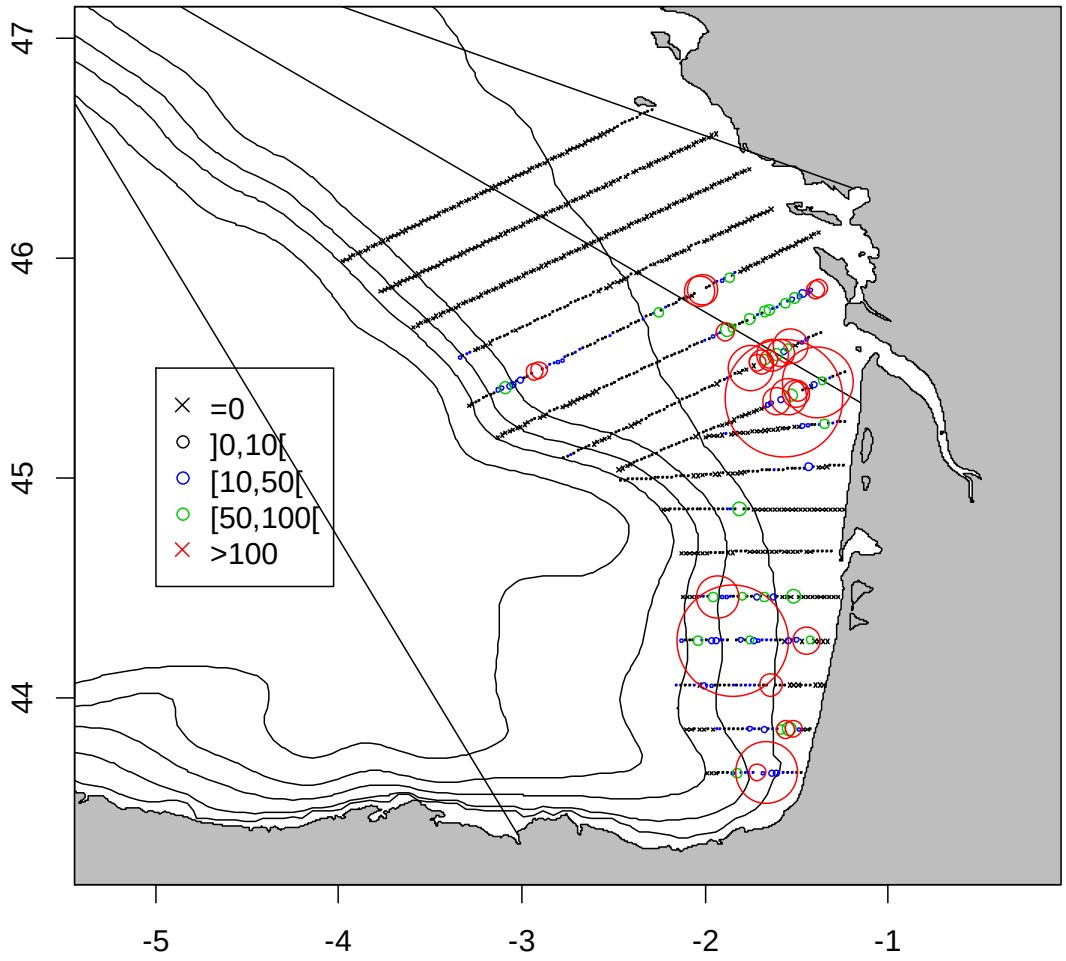


Fig. 1 Proportional representation of the anchovy concentration data (2002 Pelgas survey).

Units are tons per nautical mile square

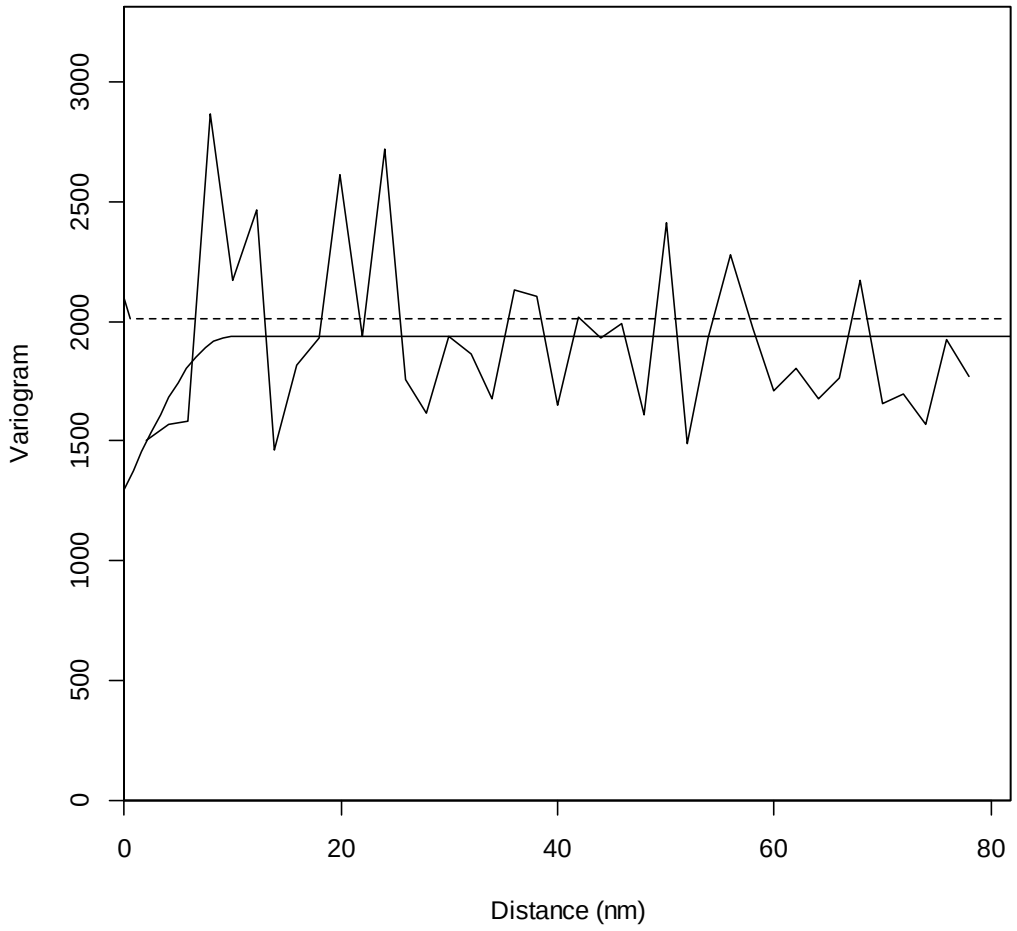


Fig. 2 Omnidirectional variogram of the target variable *Z(x)* and its model. The model is:

1265.8 Nugget + 668.2 Spherical (range=9 nm)

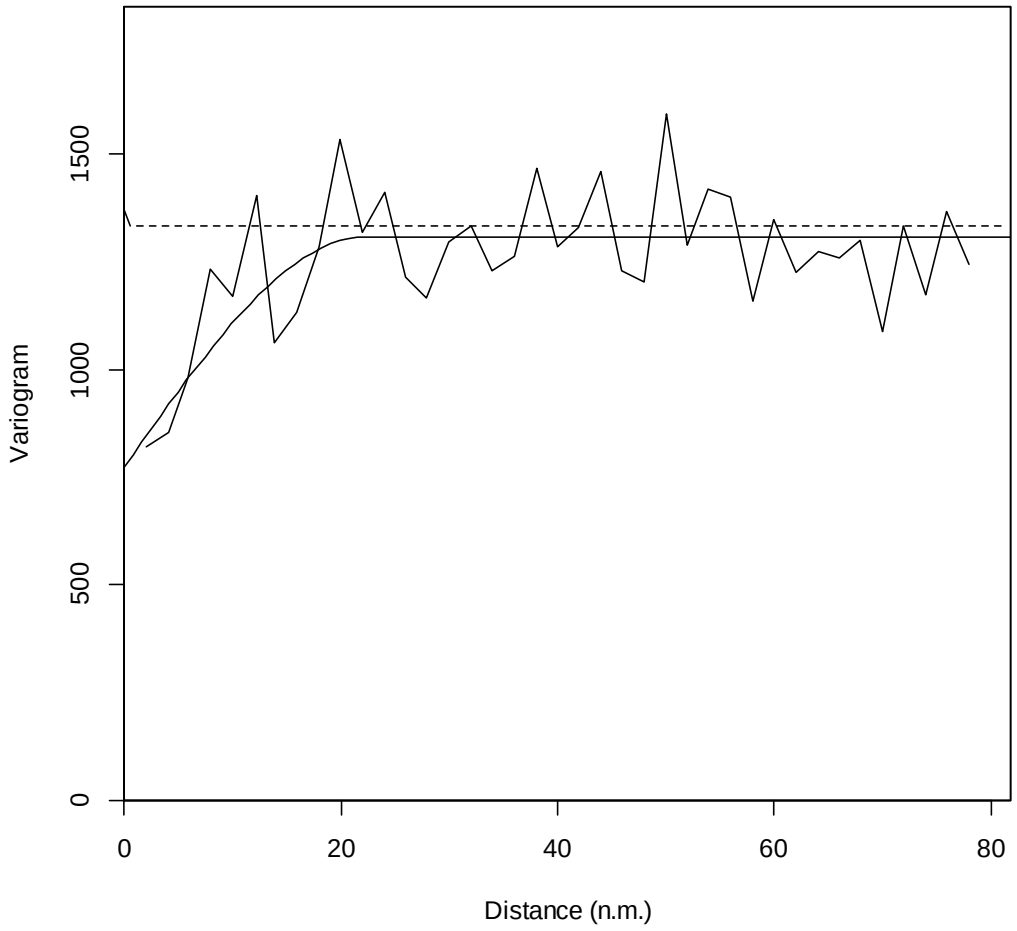


Fig. 3 Omnidirectional variogram of the discrete variable $\hat{Z}(x)$ and its model. The model is:

770 Nugget + 550 Spherical (range=22 nm)

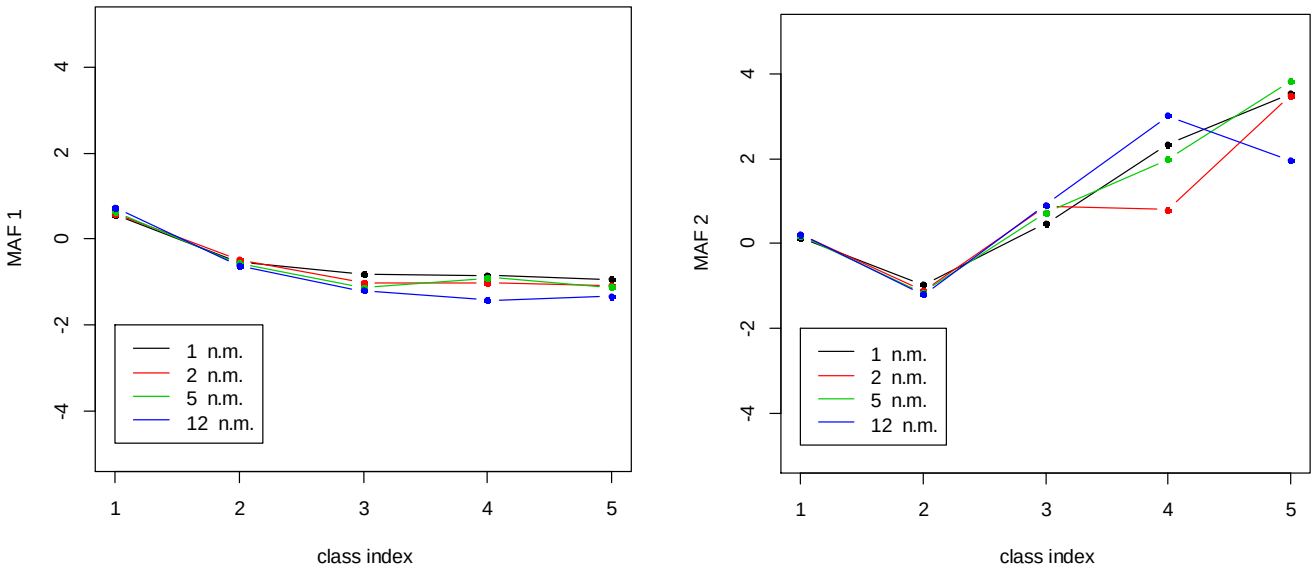


Fig. 4 Mean MAF values for the first (left) and second (right) MAF as a function of the data class index for different values of the lag Δ used to compute the MAFs. Data classes are defined in Table 1

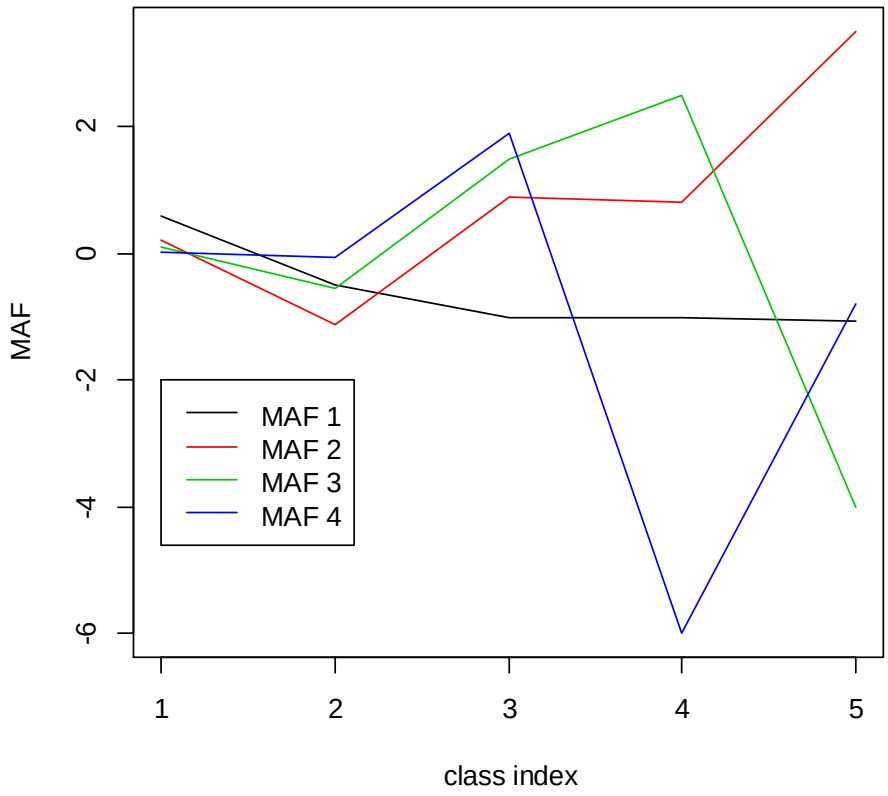


Fig. 5 Mean MAF values as a function of the data class index. MAFs are computed with lag Δ = 2nm. Data classes are defined in Table 1

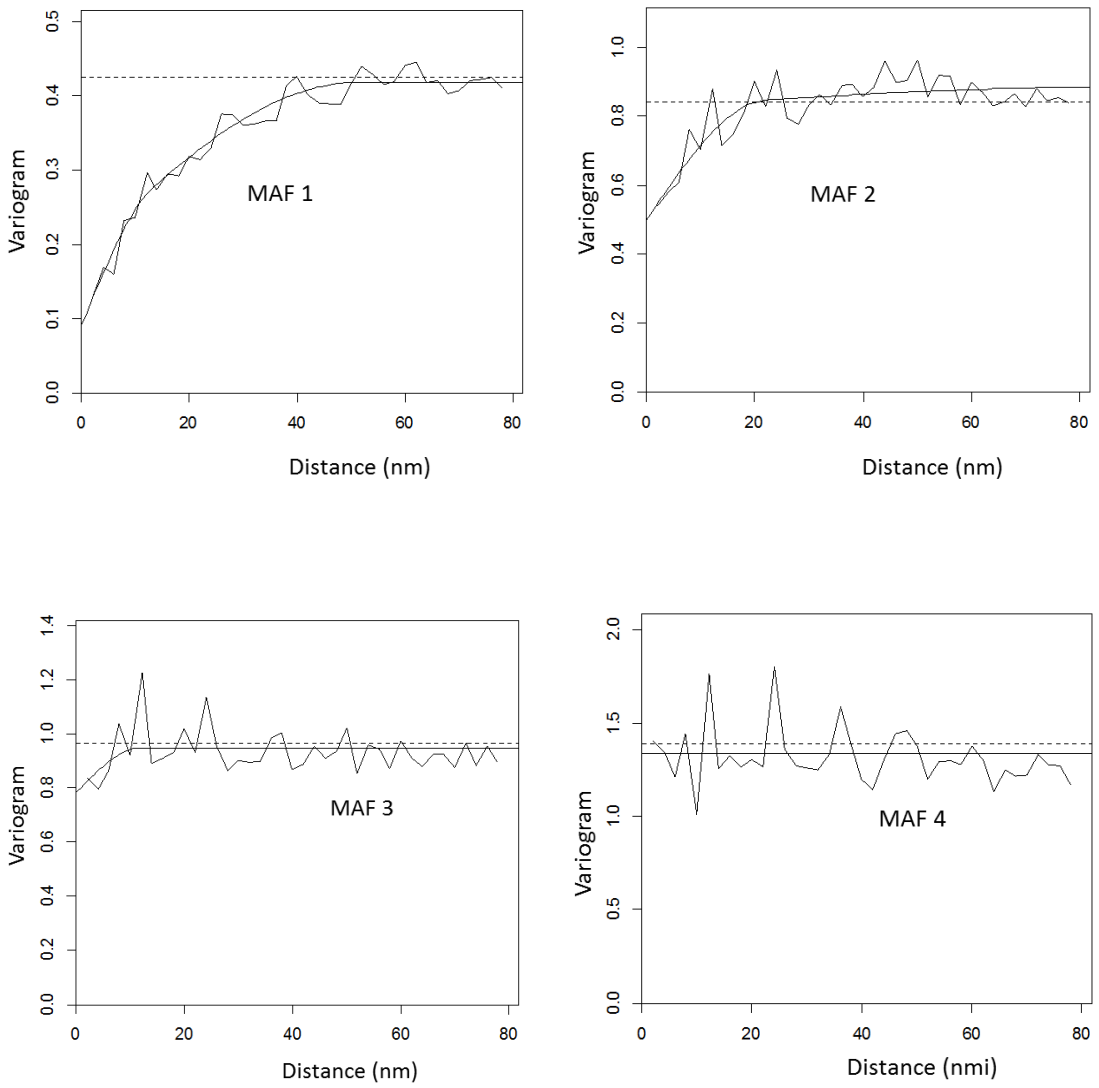


Fig. 6 Variograms of the MAFs of indicators and their models. MAF 1 model is: 0.09 Nugget + 0.10 Spherical (range=13.35) + 0.23 Spherical (range=50.74). MAF 2 model is: 0.5 Nugget + 0.23 Spherical (range=21.90) + 0.06 Spherical (range=80.60). MAF 3 model is: 0.79 Nugget + 0.16 Spherical (range=11.86). MAF 4 model is: 1.34 Nugget

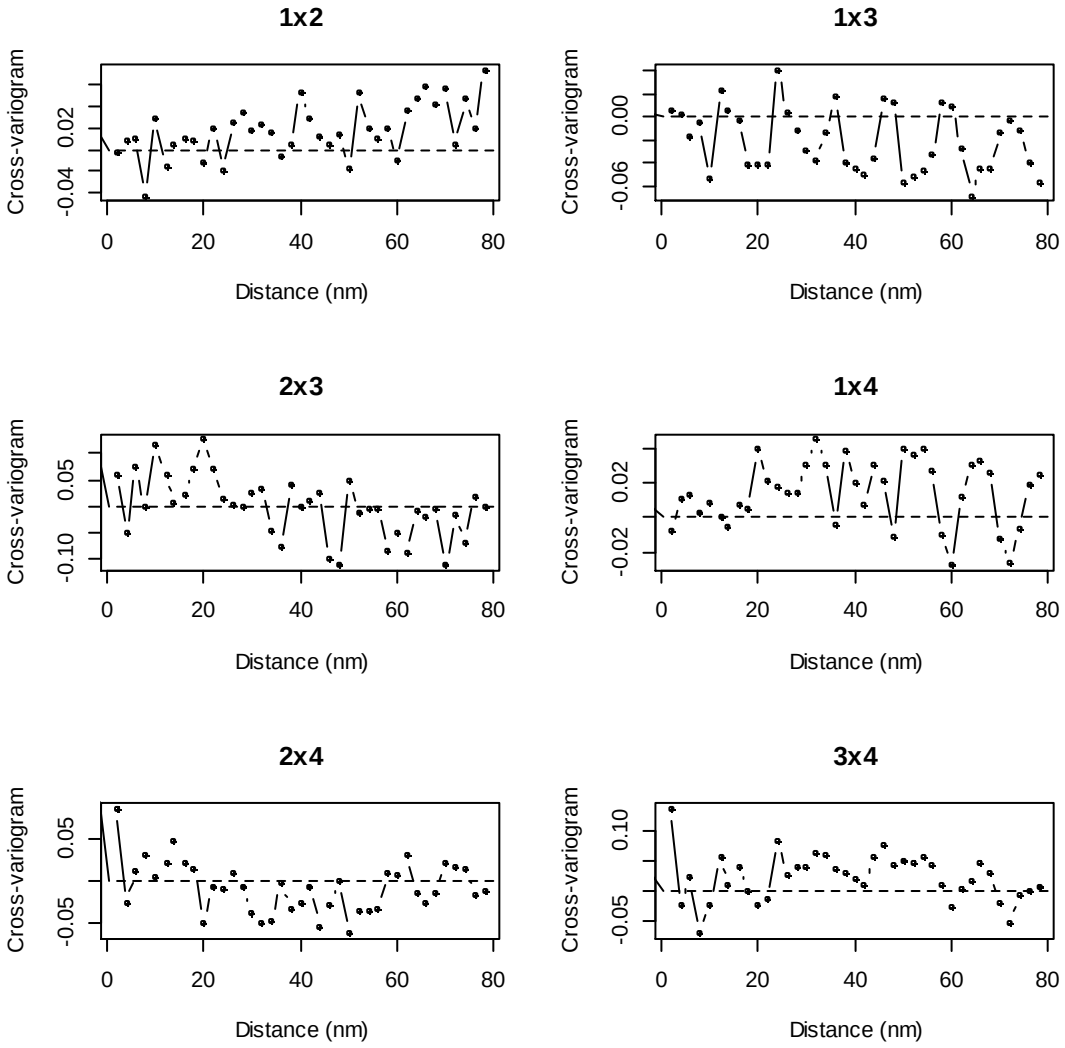


Fig. 7 Cross-variograms between the four MAFs

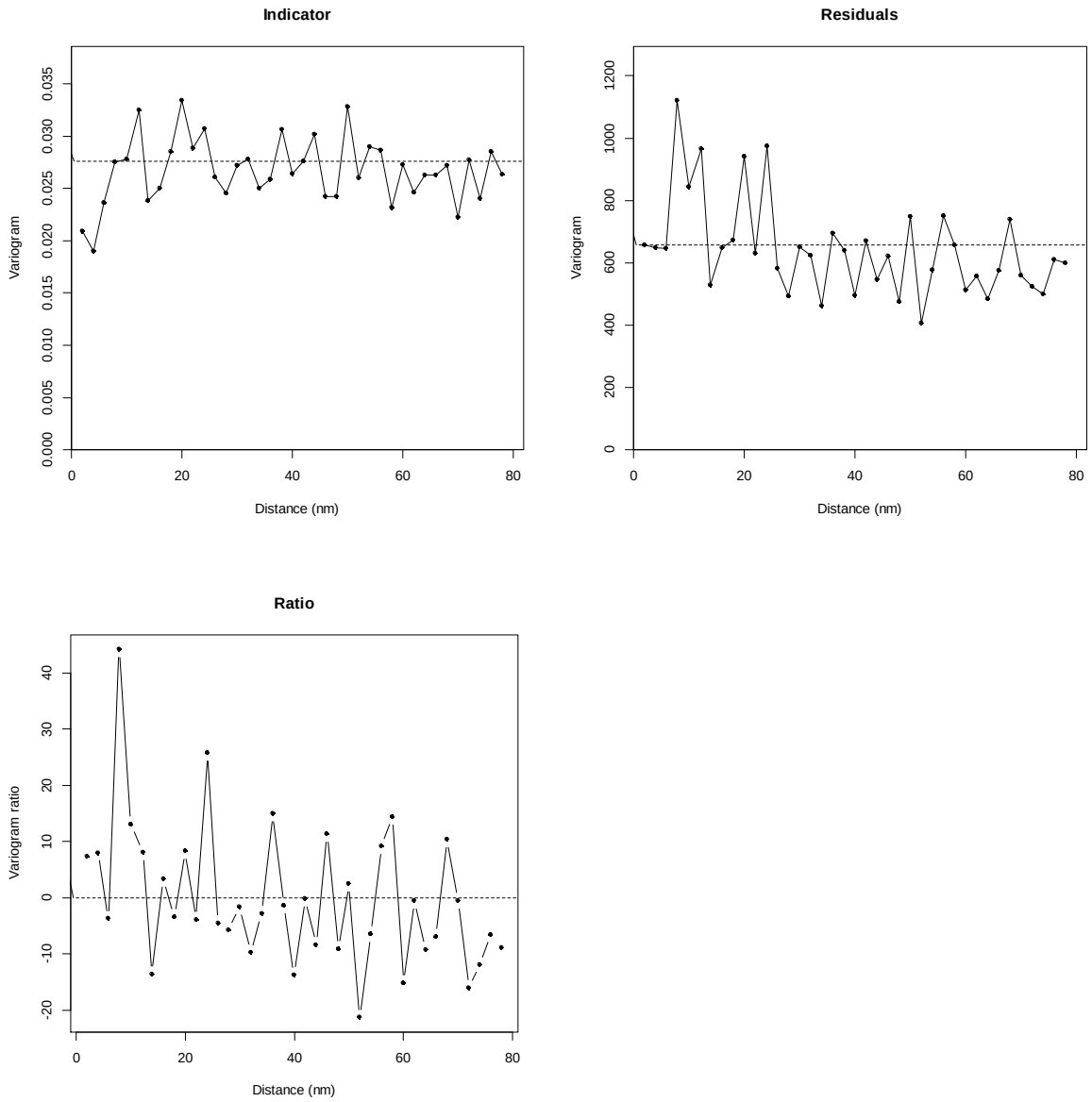


Fig. 8 Structural analysis for values higher than the top-cut cut-off. Upper left: variogram of the indicator $I_e(x) = 1\{Z(x) \geq z_e\}$. Upper right: variogram of the residual $Z_3(x)$. Lower left: ratio of the cross-variogram $I_e x Z_3$ divided by the variogram of the indicator I_e.

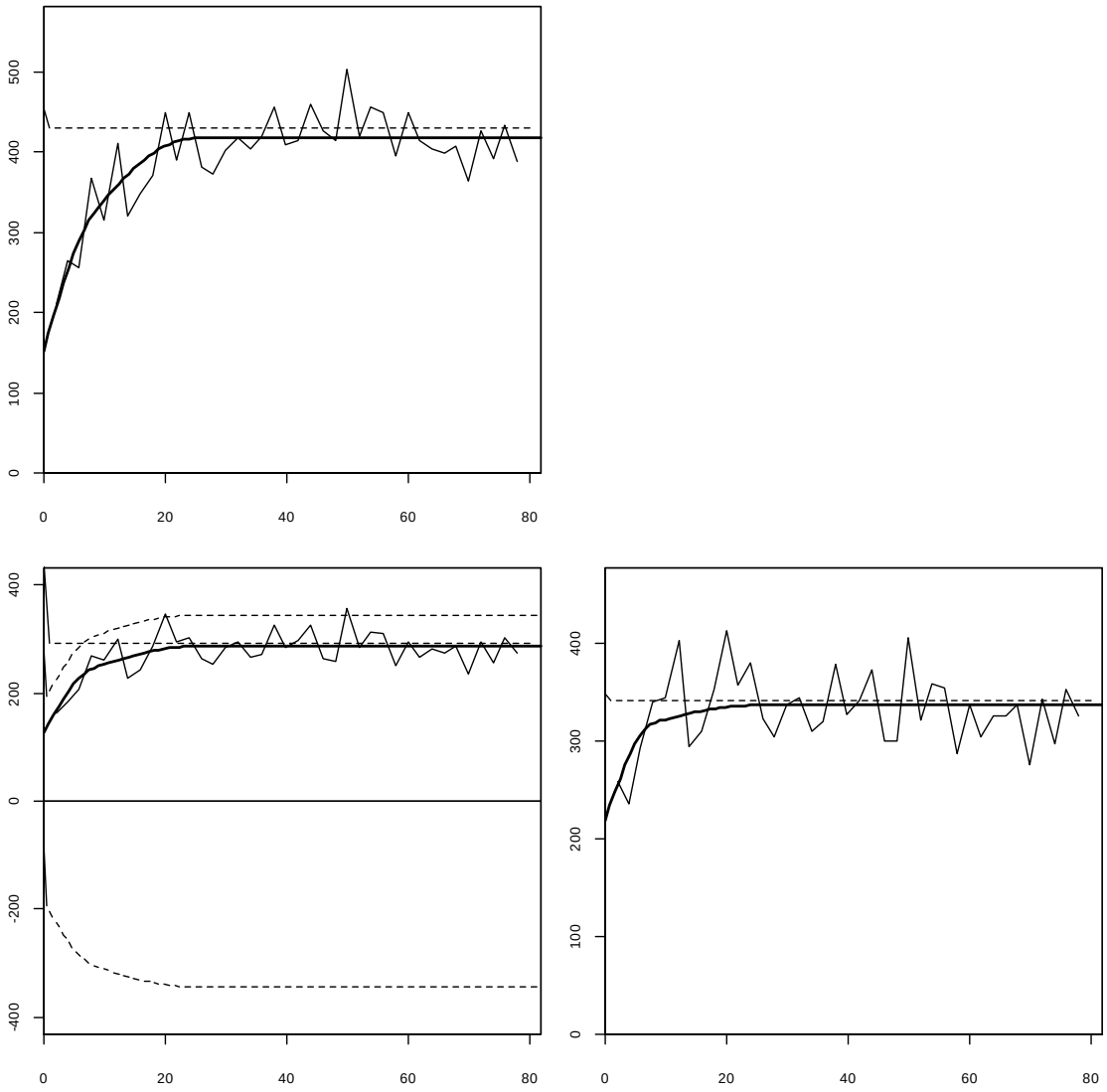


Fig. 9 Top-cut model. Upper left: variogram of the truncated variable $Z_1(x)$ and its model: 152.32 Nugget + 84.85 Spherical (range=8nm) + 180.23 Spherical (range=25nm). Lower right: variogram of the mean excess variable $Z_2(x)$ and its model: 219.36 Nugget + 84.17 Spherical (range=8nm) + 33.17 Spherical (range=25nm). Lower left: cross-variogram between $Z_1(x)$ and $Z_2(x)$ and its model: 125.79 Nugget + 83.30 Spherical (range=8nm) + 77.33 Spherical (range=25nm)

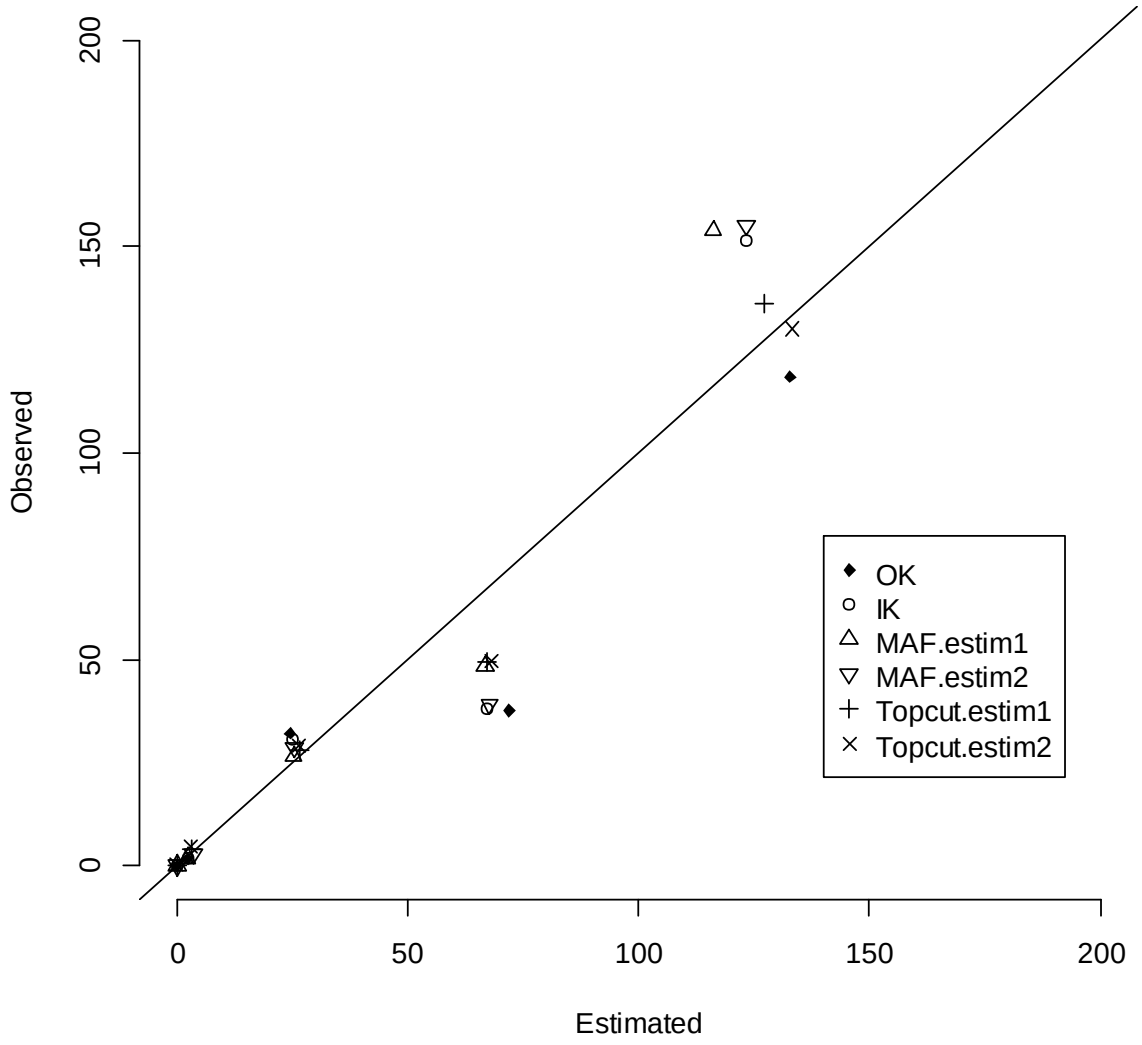
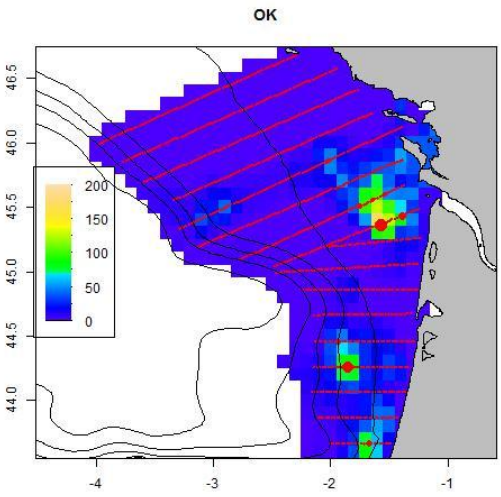
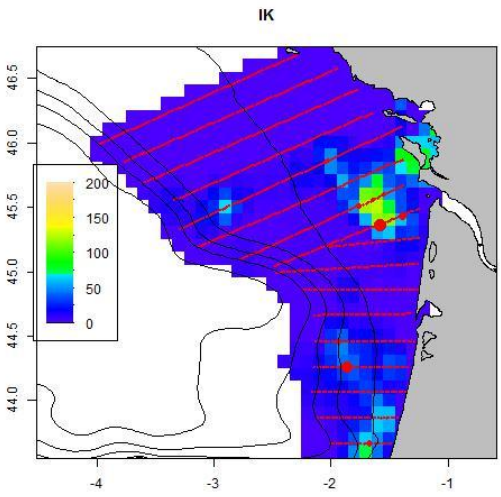
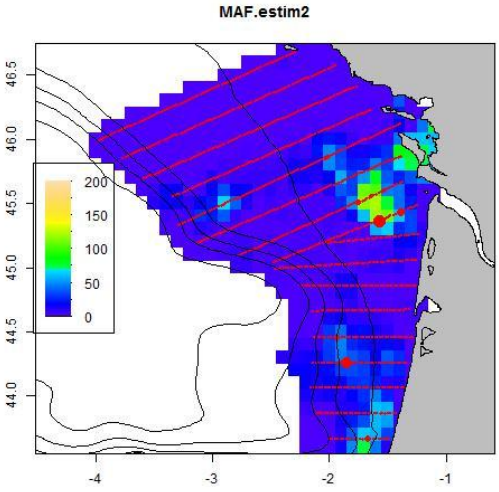
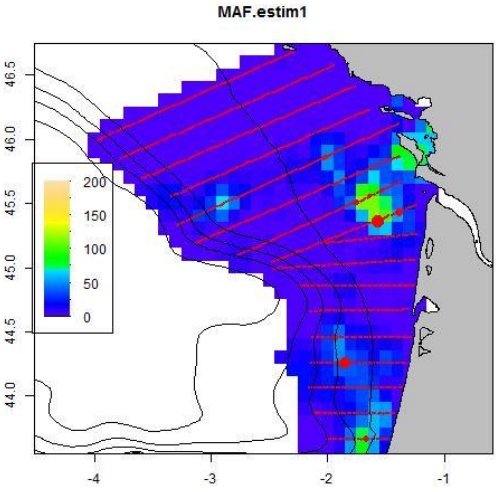


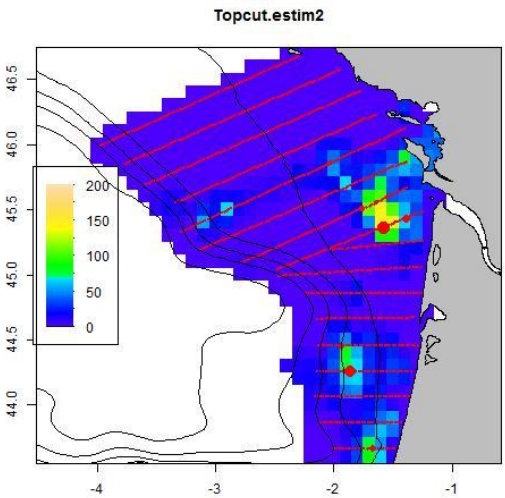
Fig. 10 Observed versus estimated mean values in each data class for the different kriging models. The line is the first bisector. The models are: ordinary kriging (OK), indicator kriging (IK), MAF model without (estim1) and with (estim2) nugget component, Top-cut model without (estim1) and with (estim2) nugget residual


OK
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1


IK
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1


MAF.estim2
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1


MAF.estim1
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1


Topcut.estim2
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1

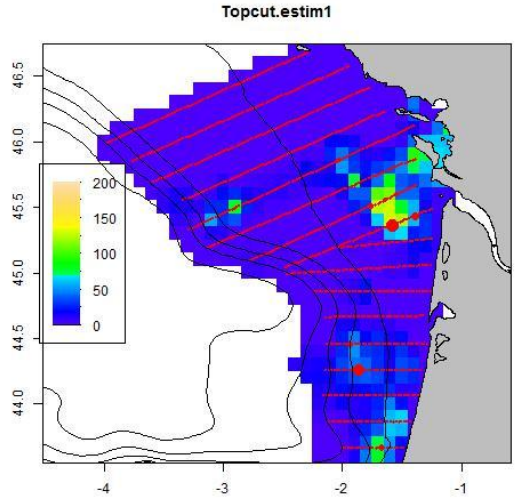

Topcut.estim1
46.5
46.0
45.5
45.0
44.5
44.0
200
150
100
50
0
-4
-3
-2
-1

Fig. 11 Maps of anchovy (tons nm^{-2}) in 2002 obtained with ordinary kriging (OK), indicator kriging (IK), and kriging with the MAF and top-cut models. For these latter models, kriging was performed without (estim1) and with (estim2) nugget component. The data are shown in red

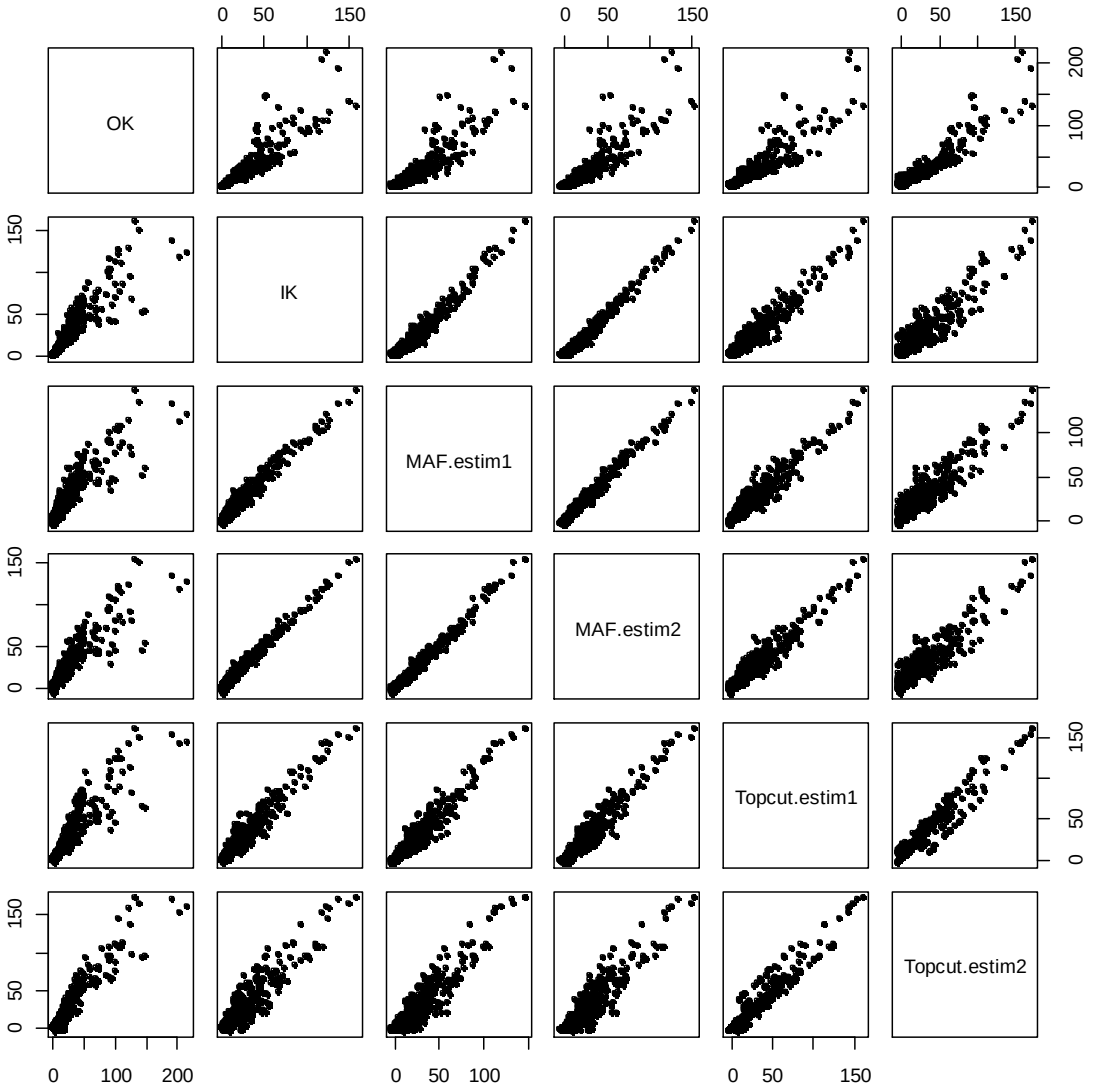


Fig. 12 Comparison of kriged map values obtained with the different models: ordinary kriging (OK), multiple indicator kriging (IK), kriging with the MAF and top-cut models. For these latter models, kriging was performed without (estim1) and with (estim2) nugget component

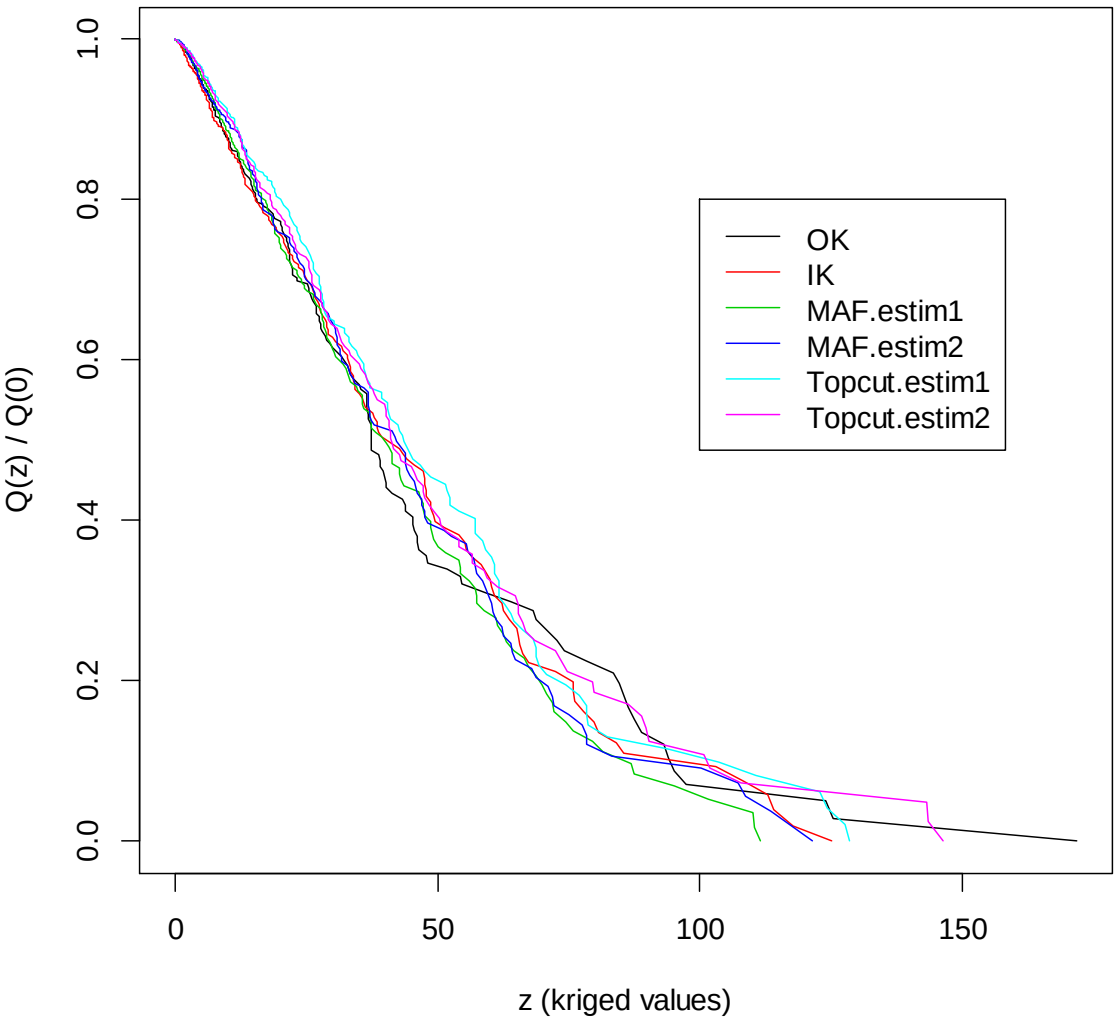


Fig. 13 Normalized selectivity curves Q(z)/ Q(0) for the kriged maps obtained with the different models: ordinary kriging (OK), indicator kriging (IK), kriging with the MAF and top-cut models. For these latter models, kriging was performed without (estim1) and with (estim2) nugget component