Using a Markov-type model to combine trawl and acoustic data in fish surveys

Mireille Bouleau, Nicolas Bez

Using a Markov-type model to combine trawl and acoustic data in fish surveys

Mireille Bouleau and Nicolas Bez

Ecole des Mines de Paris, Centre de Géostatistique, 35 Rue Saint Honoré, F-77305 Fontainebleau, France [tel: +33 1 64694778, fax: +33 1 64694705, e-mail : mireille.bouleau@ensmp.fr, nicolas.bez@ensmp.fr].

1 Introduction

Fisheries management is based on estimations of fish abundances derived from commercial catches. Models used to produce these estimates are, most of the time, tuned with indices of abundances estimated from scientific surveys. In the Barents Sea used for application in this paper, the surveys consist in deploying a net every twenty nautical miles (n.mi.). With the objective to compensate for this large distance between catches, acoustic measurements are also collected all along the vessel track when the vessel is shipping from one station to the next. This additional measure of fish concentration does not actually capture fish but estimate their number through their echoes (echoes of all the fish present in the insonified cone beneath the boat). Acoustic echoes are generally integrated over regular distance bins (say one nautical mile) and provide a spatially very dense sampling of fish distribution but different in nature from the spare tows. The purpose of the study is to take as much as possible advantage of this additional information for estimation and mapping purposes.

Here, we consider a partially heterotopic sampling where the target variable is observed on a subset of the auxiliary variable samples. Theoretically cokriging allows performing estimates in such heterotopic configurations. However it can become difficult when the number of samples is high or/and when spatial structures are difficult to model. In such cases, simplifications either assumed or data controlled, are welcome. For instance, for two variables, a Markov-type model, also called model with orthogonal residual, is a well-known simplification (Rivoirard, 2001) as one of the two variables is self-krigeable. Two kinds of Markov-type models are mentioned in literature (Schmaryan and Journel 1999): when the cross structure is proportional to the structure of the auxiliary variable or when it is proportional to the structure of the target variable. Here, we consider the first case, The trawl variable is decomposed into an acoustic and a residual components, these two components being spatially uncorrelated, but not independent. In this model, the trawl variable is subordinated to the acoustic, which is the master variable.

After a quick presentation of the data and of the notations, this paper presents the problems of the practical implementation of such a model in the particular case of strong heterotopy (hypothesis testing, structural tools, skew distributions).

2 Data and Notations

Six scientific Norwegian winter surveys (1997-2002) in Barents Sea are used. The sampling scheme (i.e. the tow locations) is targeting a regular grid with a haul every 20 n.mi (Figure 1-a). Sampling size is quite large as surveys get between 200 and 300 hauls. The mean towed distance is 1 n.mi. The acoustic data turned into Nautical Area Scattering Coefficient (NASC) and expressed in $m^2 \cdot n.mi^{-2}$ (MacLennan et al. 2002) are collected continuously along the vessel track during and between trawl hauls (Figure 1-b). In this study, acoustic echoes are integrated vertically over the first 40 meters above the bottom (this was found to provide the larger correlation between the two variables) and horizontally over fixed distance bins of 1 n.mi. Given this latter parameter, between 5000 and 7000 acoustic records are available in each survey.

To get variables with comparable units, the fish catches are turned into an equivalent acoustic energy, i.e. the acoustic energy that the fish caught in the trawl hould have generated. Because fish characteristics influence this transformation, two groups of fish have been used: demersal (bottom) fish and pelagic (mid water) fish. For each group of fish, the equivalent NASC of the corresponding fish in the net is provided. The trawl variable will refer alternatively to the demersal or the pelagic equivalent NASC depending on which of these two variables happen to get larger correlation with the acoustic variable.

We get then two measurements of fish abundance (trawl and acoustic) available at equivalent supports (1 n.m.), expressed in the same similar units but sampled differently. They are modelled by two random functions: $T(x)$ the trawl and/or the target variable available at the sampling locations, $x_\alpha \in \{stations\}$ and $A(x)$ the acoustic and/or the auxiliary variable available at the sampling locations $x_\alpha \in \{stations + underways\}$.

When sampling skew distributions, the experimental variance varies considerably with the number of samples, especially when this number is low (for a given number of samples, the sampling fluctuations of the variance are all the more important that the variance is large). We observe (Table 1) that the ratio k^2 between the variance of the underway acoustic observations (few thousands data) and that of on station observations (few hundreds data) diverges from 1. This problem is referred to hereafter as “the variance discrepancy problem”. Proportional effects are examples of this problem.

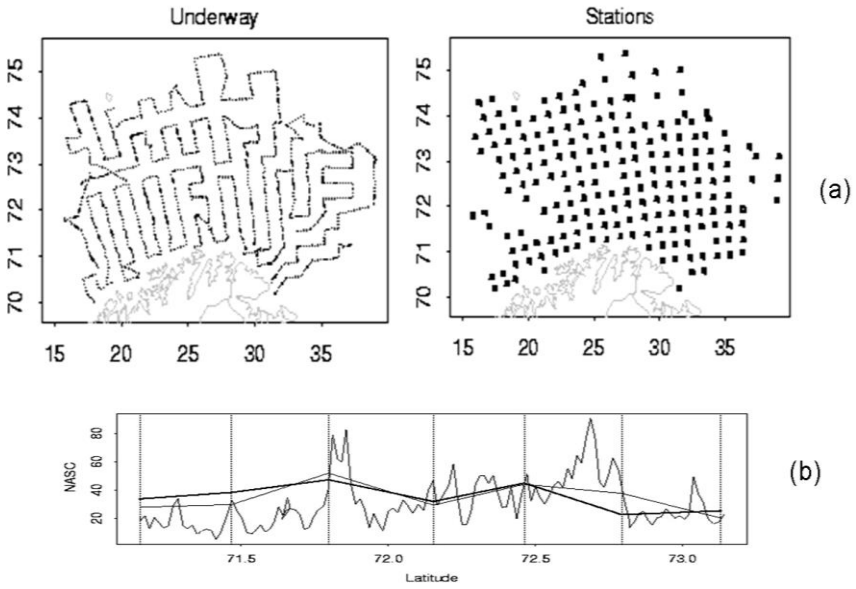

Fig. 1. (a) Locations of underway recordings (left) and of stations (right). Survey 1998. X-axes unit is in degrees of longitude and Y-axes unit is in degrees of latitude (b) Representation of a N-S section of the vessel track. The vertical dotted lines represent the stations locations. The fluctuant slight curve is the acoustic underway, the slight line joins the acoustic on-stations values and the bold line joins the demersal NASC-equivalent values collected on-stations. Distances are in degrees of latitude.

Table 1 Ratio between the variance of the underway acoustic observations and the variance of the on station acoustic observations

Year	$k^2 = \frac{\mathrm{var}\left(A(x_\alpha), \alpha \in underway\right)}{\mathrm{var}\left(A(x_\alpha), \alpha \in station\right)}$
1997	1.33
1998	1.83
1999	2.23
2000	1.35
2001	3.55
2002	2.65

Let us consider the entire line followed by the vessel during a survey. This line is made of N underway acoustic values located at the centre of their segment of 1 n.mi. each. Let us also consider a subset of the n segments, to be considered as the stations following a regular sampling with random origin (given the sampling de-

sign N = 20.n). In that case, the additive relation of the dispersion variances applies:

$$D^2(segment|line) = D^2(segment|stations) + D^2(stations|line) \quad (1)$$

The term of the left-hand side is the average variance of underway data while the first term on the right-hand side corresponds to the average variance of station data. In case of pure nugget effect, the third term equals $nugget \cdot \left(\frac{1}{n} - \frac{1}{N}\right)$ and is negligible with regards to the other terms. In this study, we have assumed that the spatial structure is short enough to neglect the dispersion variance of the stations in the line. This amounts to assume that the variances of the underway data and of the on station data are similar on average. Actual differences are then explained by the sole statistical fluctuations and are corrected for by a multiplicative term k^2 (see part 4.1 variance rescaling).

3 Methods

3.1 Model and estimation

One can show (e.g. Rivoirard, 2001) that if the acoustic is autokrigeable, its cross covariance with the trawl variable is proportional to its covariance:

$$C_{A,T}(h) = \alpha\, C_A(h) \quad (2)$$

and the trawl variable is linearly related to the acoustic up to an additive spatially orthogonal residual *R(x)*:

$$T(x) = \alpha . A(x) + \beta + R(x) \quad (3)$$

$$C_{A,R}(h) = 0 \quad \forall h \quad (4)$$

The target variable is then subordinated to the auxiliary but master. This model has a "Markov-type" property as, in Gaussian case with known means, *A(x+h)* and *T(x)* are independent when *A(x)* is given (conditional independence, Chilès and Delfiner, 1999). More generally the screen effect makes the cokriging weight of *A(x+h)* equal to zero when *A(x)* is known, whatever the histogram of the data.
The model is factorized with the two factors *A(x)* and *R(x),* and the cokriging of the target variable reduces to the sum of two krigings as the acoustic variable is known at any location where the trawl variable is known:

$$T^{CK}(x_0) = \alpha\, A^K(x_0) + \beta + R^K(x_0)$$

$$\text{where} \begin{cases} A^K(x_0) = \sum_{\substack{stations \\ + underways \\ \in \\ neighbourhood}} \lambda_\alpha^A A(x_\alpha) \\ R^K(x_0) = \sum_{\substack{stations \\ \in \\ neighbourhood}} \lambda_\alpha^R R(x_\alpha) \end{cases} \quad (5)$$

and the cokriging variance is:

$$\sigma_T^{CK}(x_0) = \alpha^2\, \sigma_A^K(x_0) + \sigma_R^K(x_0) \quad (6)$$

The constant β is in practice filtered by the ordinary kriging of the residual, and does not need to be assessed.

The estimation of the target variable at a point where the acoustic is known (underway) only uses the acoustic at the target point and on station (by the residual). Then, in the Markov-type model, cokriging is multi-collocated: for estimating an underway point, the auxiliary variable is only used at the target point and on stations. It is the only case where the cokriging is collocated (Rivoirard 2001). In a different model, the previous estimation is only an approximation of cokriging.

3.2 Practical implementation in partially heterotopic samplings

Compared to cokriging in general, an advantage of the previous estimation, based on residual, is that cross structures do not need to be modelled. Cross structures only serve to experimentally test for the validity of the model.

Two tools are used to test for the proportionality between the cross and simple structures; the cross variogram and the cross covariance. The cross variogram, not restricted to stationary cases, uses, only on station data ("isotopic tool") and misses short scale structures. The cross covariance, or preferentially in strong heterotopic cases, the cross correlogram, assumes stationarity but uses all the available information ("heterotopic tool").

The advantage of the estimation based on residual (no cross structure model) is compensated by the need to estimate the parameter α. Equation 4 is general and not specific to any sampling scheme. However, in the particular case of partially heterotopic sampling, Equation 4 is viewed as an "on station" relationship to be parameterised with on station data only and applied to underway data afterwards. In this case, rescaling is required. As a matter of fact, theoretically, the cokriging estimation variance is necessarily less or equal than the kriging estimation variance as long as the same data and the same model for the target variable are used. Here, the model is parameterised on a subset of a data which happens to be less variable (variance discrepancy problem). When applied to the more variable underway acoustic data, it does not protect from inconsistent estimation variances.

To solve this problem, the cokriging variance has to be rescaled, so that finally we have:

$$\sigma_T^{CK}(x_0) = \frac{\alpha^2}{k^2}\sigma_{A,un}^{K}(x_0) + \sigma_R^{K}(x_0) \tag{7}$$

where $\sigma_{A,un}^{K}(x_0)$ is the acoustic kriging variance, when all the data available are used, i.e., the stations and the underways.

The estimation of the parameter α can be made by many ways theoretically equivalent. It can be estimated by the slope of the linear regression of T(x) on A(x). This approach has the advantage to allow quantifying the quality of the estimation (e.g. visual inspection of the scatter plots, R-square, etc). A weakness of the regression is that only the pairs of samples at the same location contribute to the estimation. An alternative is to use the mean ratio between the cross and simple experimental variograms computed only with data on station. The gain of this approach is to take into account all distance lags. However, no quality is directly associated to the estimation of α. To enhance the robustness of the estimate, one could have used the simple variogram for all the underway observation or cross covariances. However, the advantage of using all the data is thwarted by loss of statistical coherence. We thus chose not to retain this last estimation.

4 Results

4.1 Variance rescaling

We have simulated 500 sets of 7000 lognormal data (independently) from which 500 subsets of 300 points have been taken randomly (7000 corresponds to the number of underway samples and 300 to the number of stations in 2001).

We are in a special case of pure nugget effect in the equation (1), the variance underway and on-station have to be equal in mean.

The variance and the mean of the simulated lognormal distribution are equal to the mean and the variance of the acoustic underway in 2001 (m = 63 and $\sigma^2 = 23061$). In 80% cases, the ratio k^2 between the empirical variance of the main 7000 samples and the empirical variance of the 300 subsamples is greater than 1 (Figure 2). The value 3.55 observed in 2001 (represented by a vertical dotted line) is quite singular but not impossible. When a large value is taken, the variance of the subsample becomes extreme because of the small number of samples.

So the observed discrepancy between the experimental variances can be interpreted as a sampling problem (heterotopic sampling of skew distributions) and are in no way particular to the data used in this study. In fact, it can be considered that the variance observed underway is more realistic as it is based on 20 times more

data justifying a multiplication of on-station variance. Nevertheless to be comparable to a (monovariate) kriging variance the equation (7) is based on a downscaling of the underway variance.

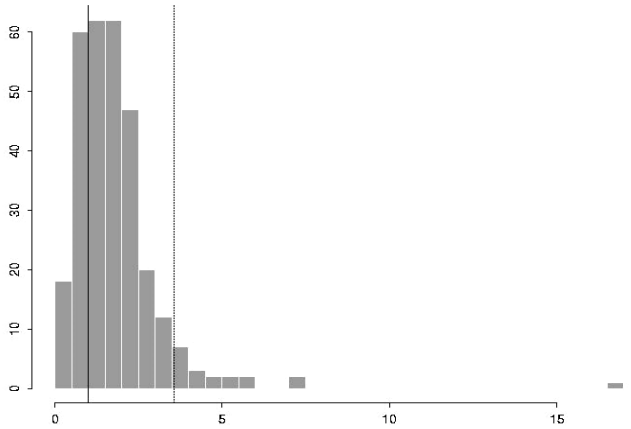

Fig. 2. Histogram of ratio between the empirical variances of the main sample (7000 points) and the subsample (300 points) for 500 draws of a lognormal distribution with the mean and the variance of the acoustic for the 2001 survey. The plain vertical line is equal to 1 and the dotted vertical line is equal to the observed ratio (3.55).

4.2 Hypothesis testing and selection of favourable cases

To test the autokrigeability assumption, experimental simple and cross correlograms have been plotted for each of the variables. Cross correlograms are potentially non symmetrical. They happened to be symmetrical and have been symmetrized before representation (Figure 3). The single and cross correlograms have been calculated along the vessel track, i.e. in one dimension: In four surveys out of six (1997-1998-2001 with demersal catches and 2000 with pelagic catches), the Markov-type model hypothesis are grounded (Figure 3, graphs with grey background). They have then been selected for application of a Markov type model.

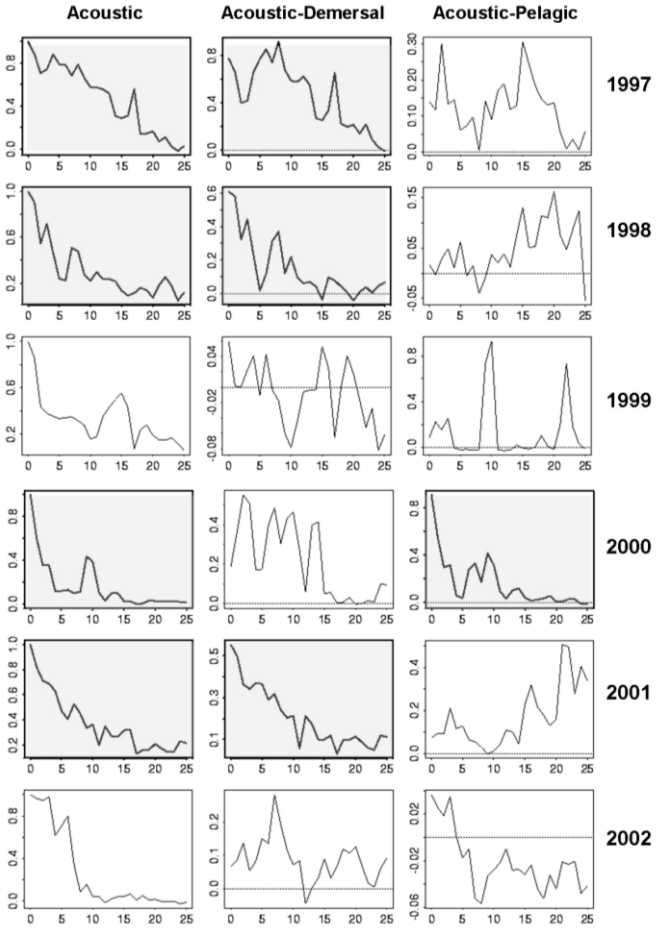

Fig. 3. Symmetrical cross correlograms calculated along the vessel track (1D). The x-axis is the distance from station (in n.mi) and the y-axis is the correlation between the acoustic underway and, according to the column: the acoustic on station (on the left), the demersal NASC-equivalent collected on station (on the middle) and the pelagic NASC-equivalent (on the right).

4.3 Estimation of parameter α

The parameter α is first estimated by the slope of the linear regression of T(x) on A(x). The cross plots between T(x) and A(x) allow evaluating visually the estimation (Figure 4). We can see that the estimations (and the R-square) are very sensitive to the large values and the fitting of the cloud is not perfect.

The parameter β (Eq. 3) of the linear regressions happens to be very small (about zero in most cases). If the additional assumption $\beta=0$ were made, the target

variable would be strictly proportional to the on-station acoustic. The α parameter would then be stable for different level of data. The estimation obtained for the whole distribution of data (Figure 4) could be processed without some outliers, or just for the low values. The estimation should probably be more robust. In fact, the estimation of α changes according to the threshold chosen and is still different for each survey.

The parameter α is also assessed by the mean ratio between the cross and simple experimental variograms computed only with data on station. The gain of this approach is to take into account all distance lags. Results obtained (Figure 5) are similar to those obtained with the regression.

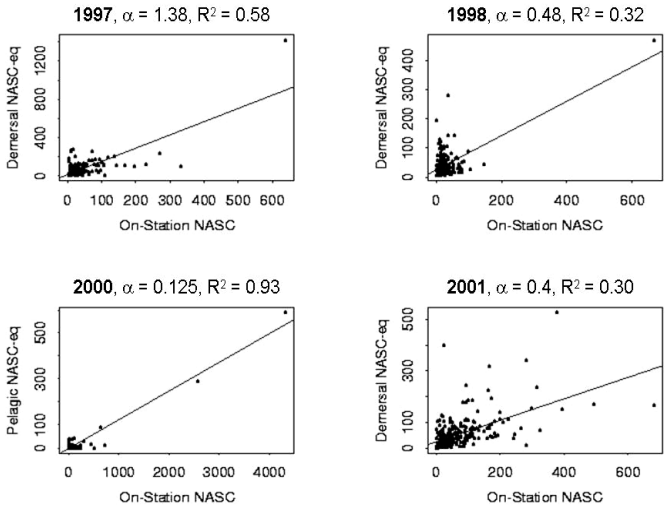

Fig. 4. Cross plot acoustic – catch on- station for the estimation of the multiplicative parameter. The lines represent the linear regressions between the two variables for each year. The values of the multiplicative coefficient α and the value of the R-square of the regression are written above each graph.

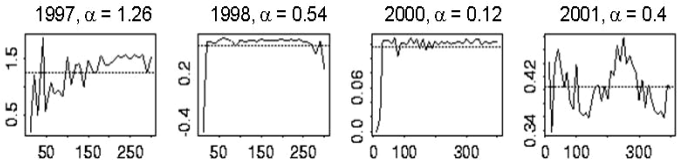

Fig. 5. Estimation of parameter α by the ratio between the $\gamma_{A,T}(h)$ and $\gamma_{A}(h)$ for on-station data only. The horizontal lines represent the mean value, i.e. the estimation. The x-axis represents distance in n.mi.

4.4 Estimations maps

To evaluate the improvement provided by the acoustic information, the bivariate approach (using a model with acoustic as master variable) is compared with a mono-variate approach based on the sole trawl values. This comparison is meaningful only if the variogram model for the trawl variable is the same for the kriging and the cokriging. It is then totally determined by the models chosen for the acoustic and the residual with the relation:

$$\gamma_T^K(h) = \gamma_T^{CK}(h) = \alpha^2\ \gamma_A(h) + \gamma_R(h) \tag{1}$$

Given that the sampling grid covers regularly the study area, the cokriged and the kriged maps have the same general long distance patterns and the use of the acoustic variable only impacts the short scale features of the distribution (Figure 6). In 1997, the kriging interpolation in the south western area where no sample is available amounts to the local mean concentration. A bivariate approach makes it possible to use the underway observations and to suggest some spatial pattern for the fish concentration in this area. In 2000, even if the weight of the acoustic is low ($\alpha = 0.12$), the cokriged map computed with a Markov-type model honours some rich areas (in the North-East) which are not observed in the kriged map of the trawl data.

4.5 Variance of the estimation error map

For the four surveys, the estimation variance is smaller for the Markov-type approach than for the single variable approach. It is not surprising since the variance of cokriging is always less or equal than the variance of the correspondent kriging.

4.6 Cross-validations

The cross validation consists in re-estimating a known point. Here we re-estimate each on-station point where the two variables, acoustic and catch, have been removed. It allows appraising the robustness of the model. For each survey the results provided are better in the bi-variate model than with the single variable model. The correlation coefficients between the estimated and observed catch values are shown in the table 2.

Table 2 Correlation coefficient between estimated and observed catch values

Year	Bivariate model	Monovariate model
1997	0.51	0.17
1998	0.30	0.09
2000	0.39	0.06
2001	0.41	0.34

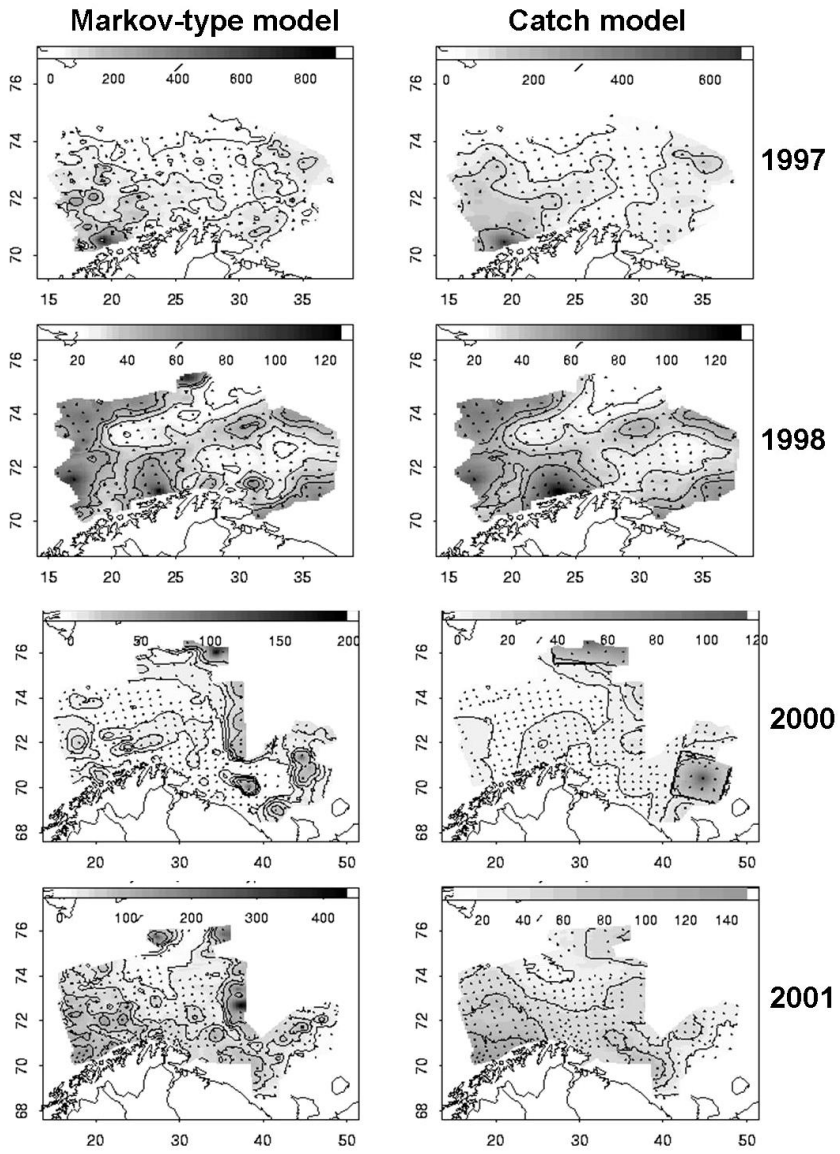

Fig. 6. Estimation maps obtained by the Markov-type model (left column) and a simple model using only the catch information available on station in a compatible model (right column). The maps on the left hand side are very more detailed. To compare the models, the grey scales are identical for each year but different from survey to survey.

5 Discussion

The estimation of the α parameter is a key step of the process as it this parameter quantifies the weight of the acoustic. In the model, the acoustic drives the catches and the residual allows rescaling the estimation on stations. Such behaviour is physically well understandable: acoustic provides a good representation of the fish abundance and the fish abundance is just obtained by adding a corrective term calculated by the divergence observed on stations between acoustic and catches (the residual). The main structure is then provided by the acoustic and the residual, in the general case, would not be strongly structured. However, in practice, the residual can have a long range structure because of one or few large values at the edge of the sampling area.

The use of an auxiliary variable largely more densely sampled than the target variable improves its estimation. The bi-variate model improves the estimation of the catch by combining acoustic with a simple relation exhibiting the role of each variable. However it is important to mitigate the results at least by the quality of the estimation of the parameter α. This key parameter has to be estimated and the quality of its estimation drives the quality of the whole process. When variables get skew distributions like in the present study, once again, linear approaches happen to be fragile and we have indeed a weak confidence in the actual value of this parameter.

When an estimation routine need to be processed every year, like the estimation of fish abundance, it is important to find a model robust enough to work for all the configurations, not only for a particular year with particular relation between the variables. Here the assumptions of the regression model are funded in four surveys out of six. For the two other cases (1999 and 2002) the erratic cross-structures do not allow to conclude to any model. The fact that the catch is driven by the acoustic, can be considered like a physical property, and we can think that the model will be also pertinent for the next years.

Because of the large skewness of the data, the use of linear approaches is questionable. Linear tools are indeed very sensitive to the large values which often hide the behaviour of the lower values (Rivoirard et al. 2000). Some non-linear tools like disjunctive kriging allow minimizing this impact. However the computation of a bivariate disjunctive model is laborious and requires heavy assumptions (Goovaerts 1997). The leading idea of this study has been to find a model simple enough and robust enough to be relevant in most available surveys.

Acknowledgement

The authors thank the Institute of Marine Research of Bergen who carried out all the surveys used in this study, in particular Olav Rune Godø and Vidar Hjellvik for formatting the databases and for helping our understanding of the Barents Sea features.

References

Chiles, J-P., Delphiner, P. (1999) Geostatistics, Modelling Spatial Uncertainty, Wiley, New York, 695p.

Goovaerts P. (1997) Geostatistics for Natural Resources Evaluation , Oxford Univ. Press.

MacLennan, D.N., Fernandes, P.G. and Dalen, J. (2002) A consistent approach to definitions and symbols in fisheries acoustic. ICES Journal of Marine Science, 59: 365-369.

Rivoirard, J. (2001) Which models for collocated cokriging? Math. Geol, v.33, no 2, p117-131.

Rivoirard J., J. Simmonds, K.G. Foote, P. Fernandez and N. Bez (2000) Geostatistics for Estimating Fish Abundance. Ed. Blackwell Science, 206p.

Schmaryan L. and Journel A.G (1999) Two Markov-type models and their application, Math. Geol. Vol. 31, no 8, pp 965-98