International Council for
the Exploration of the Sea
ICES / CIEM

CM 2001/P:17

Allocation of survey effort between small scale and large scale and precision of fisheries survey-based abundance estimates

by

P. Petitgas

Fisheries surveys are undertaken with the primary objective of monitoring the abundance of fish stocks. Their design is usually large scale and made to cover large areas of sea allowing the estimation of regional average values of fish density. Therefore the design does not allow for intensive local sampling. But the local and meso-scale aggregation of fish is generally an important feature of the fish spatial distribution representing about half of the total variability in the spatial distribution. The local heterogeneity is in general not well sampled by the large scale fisheries surveys. Yet, it contributes significantly to the error variance on the global abundance estimate. The present study analyses the intensity with which local fish aggregations should be sampled for improving the global abundance estimate. The International Bottom Trawl Surveys (IBTS) and the German Small Scale Surveys (GSSS) in the North Sea are jointly analysed and taken as examples. The IBTS (i.e., large scale) survey strategy is to take one or two samples in each ICES statistical rectangle. The GSSS (i.e., small scale) survey strategy is to sample intensively during a few days in a chosen rectangle (i.e., box). The small scale surveys give knowledge on the level of variance within the boxes. The variance in the IBTS data is thus generated by a mixture of the boxes variances. The coherence in the level of variances between IBTS survey and box surveys is analysed for cod age 2 in quarter 2 using geostatistics. It is shown that the small scale box surveys contain as much variance as the large scale IBTS survey. The question is raised whether the large scale IBTS survey under-estimates the overall variance over the North Sea. To analyse this question simple theoretical fish distributions were simulated and then sampled with different allocation of effort between small scale and large scale. Results show that allocating more sampling effort at small scale leads to over-estimate the overall variance as well as the small scale variance term. This study is a contribution of the EU-funded project FINE.

Keywords: Survey Design, Geostatistics, Variance, IBTS.

P. Petitgas: IFREMER, BP 21105, F- 44311, Nantes, France [tel: +33 240 374000, fax: +33 240 374075, e-mail: Pierre.Petitgas@ifremer.fr].

1. A geostatistical comparison in the level of variance in the large scale and small scale surveys : North Sea cod age 2 in quarter 2 as an example.

The spatial unit division of the North Sea is the ICES statistical rectangle. The IBTS (ie, large scale) survey strategy is to take one or two samples in each rectangle. The small scale survey strategy is to sample intensively during a few days in a chosen rectangle (ie, boxe). The small scale surveys give knowledge on the level of variance within the boxes. The variance in the IBTS data is thus generated by a mixture of the boxes variances. The coherence in the level of variances between IBTS survey and boxe surveys was analysed for cod age 2 in quarter 2. The interest in such study is the understanding of how small scale variability influences the large scale survey data and subsequently analyse if IBTS survey strategy is adequate.

The data : maps and basic statistics

The data are fish numbers per hour of trawl haul. Fig. 1 shows a map of the IBTS data with localisation of the boxes where intentsive sampling has been performed. We see an overall pattern for cod age 2 distribution : more fish in the center of the North Sea, ie a large scale patch. The boxes are positioned in high density areas and on the border of the large scale patch. Fig. 2 shows the maps for the different boxes. Table 1 gives basic statistics for IBTS and boxes data. The analysis of a particular box is not of much interest as it is considered one micro scale view out of many. We are interested in the average box parameters and their relation with the large scale ones. Average statistics for all boxes have therefore been computed (Table 1).

Data structure : histograms and variograms

Fig. 3a shows the histogram of the IBTS values. It has many zeroes and is skewed to the right. The histogram tail is not well known and influences precision on the mean estimate. This is a typical situation with survey data. Fig. 3b shows the histogram for the box data pooled. It is less skewed than the IBTS data and contains less zeroes. Fig. 4 shows a qq- plot (quantile to quantile correspondence) between large scale values and pooled values for the boxes data. Values have been centered to their means in each box. We see good correspondence except for low and high values. It is thought logical that the IBTS contains more lower residuals than boxes because boxes have not been chosen in low value areas. It is interesting to see that IBTS shows higher residuals than the boxes. This is can be due to (i) a lower IBTS mean, (ii) the fact that by chance, IBTS has sampled a high density which small scale surveys have not, (iii) the fact that no boxe is positioned inside the large scale patch.

Fig. 5 shows the variogram of the IBTS data. The model fitted is the sum of a nugget and 2 spherical models (Table 2). The model was fitted so as to follow closely the experimental points and to reproduce the overall level of the data variance (see next section). The long range structure can be attributed to the large scale patch. Fig. 6 gives variograms in each box. Structure is difficult to evidence due to the high level of variance and to the small nb of points. No variogram models were fitted.

Estimation of variances

The variogram enables to compute model based estimates of variance at any scale. Average variance within a polygone v equals the average of the variogram for all distances in v (Matheron 1971):

$$\bar{\gamma}_{vv} = \frac{1}{v^2} \int_v \int_v \gamma(|x-y|) dxdy .$$

This is the dispersion variance over v. Its estimation was performed using the software EVA dedicated for computing variances based on the variogram (Petitgas and Prampart 1997). The IBTS variogram model was chosen so that (i) the model would fit the experimental points and (ii) so that the $\bar{\gamma}_{VV}$ would be of the same order of magnitude than the data variance (similar overall variance in the model than in the data).

Let v denote a boxe and V the North Sea. Let $D^2(./V)$ denote the dispersion variance in V (ie, the variance between point samples), $D^2(./v)$ the dispersion variance in v (ie, the average of the variances for many boxes), and $D^2(v/V)$ the variance between boxe means. These variances are related by the additive relation (Matheron 1971) :

$$D^2(./V) = D^2 (v/V) + D^2 (./v)$$

$D^2(./V)$ is estimated by the data variance or the $\bar{\gamma}_{VV}$ using the IBTS variogram model. These values are similar.

$$D^2 \text{ (./V) ibts model} = \bar{\gamma}_{VV}{}^{ibts} = 80.81$$
$$D^2 \text{ (./V) ibts data} = \text{Var[ibts data]} = 80.35$$

D^2 $(./v)$ can be estimated in two ways : (1) using the IBTS variogram model and computing the $\bar{\gamma}_{vv}$ for the boxe dimension and (2) by computing the average of the box variances.

$$D^2 \text{ (./v) ibts} = \bar{\gamma}_{vv}{}^{ibts} = 44.62$$
$$D^2 \text{ (./v) boxe} = E[\text{boxe.var}] = 114.6$$

The IBTS variogram model underestimates largely the experimental variance within the boxes, in comparison to the box surveys.

D^2 (v/V) can be estimated (1) by using the additive formula of the variances or (2) by taking the variance between the box averages.

$$D^2 \text{ (v/V) ibts} = 80.81 - 44.62 = 36.19$$
$$D^2 \text{ (v/V) boxe} = \text{Var[boxe.mean]} = 31.68$$

Estimates of variance between box means are compatible between IBTS and Box surveys. But estimates of variance within boxes are not compatible between IBTS and Box surveys. This poses the acute question of how reliable is the precision of IBTS estimates and in particular how reliable is the estimate of precision that can be computed from the IBTS data. Is the IBTS sampling strategy enable to capture all the variance?

When using the simple average for the mean estimate, the error variance of the mean estimate is estimated using the now classical geostatistical estimation variance formula (Matheron 1971) :

$$\sigma_E{}^2 = 2\bar{\gamma}_{SV} - \bar{\gamma}_{VV} - \bar{\gamma}_{SS}$$

where *S* denotes the set of sample locations and *V* the field sampled (ie, North Sea). A polygon was defined as the contour of the data samples. The average variogram values were computed using the software EVA. Results are :

$$m=5.55$$

$$\sigma_E / m = 11.3\%$$

A relative estimation error of 11.3% is to be regarded as a satisfactory survey precision. But it is difficult to be sure that this value is reliable as we have seen previously that the IBTS sampling strategy may not capture all the variance. Simulations would enable to analyse how IBTS survey strategy mixes the small scale boxes variances.

2. Random function simulations and re-sampling: Mean and variance estimates for different allocation of sampling effort between large scale and small scale.

The estimation of the survey precision relies on the estimate of variogram and variance derived from the data. Given a spatial distribution in which there would be a large scale spatial pattern and a high small scale variability, which would be the sampling scheme that would allow to estimate the true level of variance and which would be the sampling schemes that would not? Simple theoretical spatial distributions were simulated to answer this question in a heuristic manner.

The spatial distribution model considered is that of a random function which has two structural components: a large scale drift (i.e., trend) and purely random residuals. Four different scenarios were considered with an additive or a multiplicative model with the error structure of the residuals being symmetrical (e.g., gaussian) or skewed to the right (e.g., log normal). The contribution of the residual variability in the overall variance is also a parameter that could be made to vary between scenarios. The case of 50% variance in the residuals was only considered here. The models were simulated in one dimension (i.e., a segment on the line). The trend considered was linear.

Formulas for the models considered

The additive model writes: $Z(x) = m(x) + \sigma\varepsilon(x)$

The multiplicative models writes: $Z(x) = m(x) + m(x)\sigma\varepsilon(x)$

Residuals *ε(x)* have zero mean and variance one, are independent of *m(x)* and have no spatial structure (nugget effect). *Z* takes its values along the line and is analysed on the segment [0,L]. The drift *m(x)* is linear: $m(x) = b - ax$

The mean of Z over the segment [0,L] (i.e., the spatial mean) writes for one realisation: $Z_L = \frac{1}{L}\int_0^L Z(x)dx$.

Its expectation for all realisations of Z is: $M_L = E[Z_L] = \frac{1}{L}\int_0^L m(x)dx = b - \frac{a}{2}L$.

The variance of Z over the segment [0,L] (i.e., the spatial variance) writes for one realisation: $V_L = \frac{1}{L}\int_0^L (Z(x) - Z_L)^2\, dx$

Its expectation for all realisations of Z is: $D^2{}_L = E[V_L] = \frac{1}{L}\int_0^L E(Z(x) - Z_L)^2 dx$.

In the case of the additive model, $D^2{}_L$ develops as:

$$D^2{}_L = \frac{1}{L}\int_0^L E[(m(x) - Z_L)^2]dx + \sigma^2 = D^2(m(x)) + \sigma^2$$

In the case of the multiplicative model, V_L develops as: $D^2{}_L = D^2(m(x)) + \sigma^2 I$

Where I is the following integral: $I = \frac{1}{L}\int_0^L m(x)^2 dx$

The trend *m(x)* being linear with slope *a*, *D^2(m(x))* develops as: $D^2(m(x)) = \frac{a^2}{12}L^2$

And the integral I develops as: $I = b^2 + \frac{a}{3}L^2 - abL$

Similarly, the variogram of Z develops as follows. In the case of the additive model the variogram of Z writes (δ is the Dirac function modelling the nugget effect):

$$\gamma_Z(h) = \gamma_m(h) + \sigma^2\gamma_\varepsilon(h) = \frac{a^2}{2}h^2 + \sigma^2\delta(h)$$

In the case of the multiplicative model, the variogram of Z writes:

$$\gamma_Z(h) = \gamma_m(h) + \sigma^2 I\gamma_\varepsilon(h) = \frac{a^2}{2}h^2 + \sigma^2 I\delta(h)$$

From one realisation (i.e., using the survey data), one estimates the variogram of Z which is then used to estimate the survey precision. Thus, it is of prior importance that the survey design used allows a correct estimation of $D^2{}_L$, *D^2(m(x))*, *I* and *a*. In these quantities, the spatial variation of the trend *m(x)* plays an important role.

Simulations and results of sampling designs

Four models were considered: additive with normal residuals, additive with lognormal residuals, multiplicative with normal residuals *ε*, multiplicative with lognormal residuals *ε*. The variance σ^2 was adjusted to be the percent *α* of V_L, the average variance of *Z* over *L*: $\sigma^2 = \alpha D^2{}_L$.

Thus, in the case of the additive model: $\sigma^2 = \frac{\alpha}{1-\alpha}D^2(m(x))$,

and in the case of the multiplicative model: $\sigma^2 = \frac{\alpha}{1-\alpha}\frac{D^2(m(x))}{I}$.

These quantities were calculated using parameters *L*, *b*, and *a* which were fixed for all simulations. Lognormal residuals with zero mean and unit variance were obtained by generating normal values, u, from a Gaussian(0,1) and applying the following formula: $\varepsilon = \frac{e^u - e^{0.5}}{\sqrt{e(e-1)}}$.

Simulation parameters were: *L=100, b=100, a=-1, α=0.5.*
Thus: M_L*=50, I=3333.33,* D^2*(m(x))=833.33* and $D^2{}_L$*=1666.67.*

The sampling effort considered was N_x samples regularly spaced in the segment *[0,L]* with inter-sample distance being L/N_x. At each sampling location, N_e samples were taken from the probability distribution of *ε*. The sampled simulated values of *Z* are noted z_{ij} where *i* is the index of the locations x_i along the segment *[0,L]* and *j* is the index of the replicated samples at location x_i. The simple average of the N_e replicated samples at location x_i is denoted $\bar{z}_i$. For one simulation of the random function and one survey with Nx.Ne samples, the mean of variance of Z over L were estimated ($\hat{Z}_L$ and $\hat{V}_L$) as well as the residual variance and the variance in the trend ($\hat{D}^2(m(x))$ and $est(\sigma^2 I)$):

$$\hat{Z}_L = \frac{1}{N_x N_e}\sum_{i=1}^{N_x}\sum_{j=1}^{N_e} z_{ij} \;;\; \hat{V}_L = \frac{1}{N_x N_e - 1}\sum_{i=1}^{N_x}\sum_{j=1}^{N_e} (z_{ij} - \hat{Z}_L)^2$$

$$\hat{D}^2(m(x)) = \frac{1}{N_x - 1}\sum_{i=1}^{N_x} (\bar{z}_i - \bar{Z}_L)^2 ;\; est(\sigma^2 I) = \frac{1}{N_x(N_e - 1)}\sum_{i=1}^{N_x}\sum_{j=1}^{N_e} (z_{ij} - \bar{z}_i)^2$$

This was repeated $N_{sim}=100$ times (i.e., realisations) leading to N_{sim} values of $\hat{Z}_L$, $\hat{V}_L$,$\hat{D}^2(m(x))$ and $est(\sigma^2 I)$. The average and variance of $\hat{Z}_L$, $\hat{V}_L$,$\hat{D}^2(m(x))$ and $est(\sigma^2 I)$ were then computed over the N_{sim} realisations. Results are presented as follows:
Ratio of the mean estimates to the process expectations:
$m.m(Z) = E(\hat{Z}_L)/M_L$; $m.v(Z) = E(\hat{V}_L)/D^2{}_L$;
$m.v(trend) = E(\hat{D}^2(m(x)))/D^2(m(x))$; $m.v(res) = E(est(\sigma^2 I))/\sigma^2 I$
Ratio of the standard deviation of the estimates to the process expectations:
$cv.m(Z) = \sqrt{V(\hat{Z}_L)}/M_L$; $cv.v(Z) = \sqrt{V(\hat{V}_L)}/D^2{}_L$;
$cv.v(trend) = \sqrt{V(\hat{D}^2(m(x)))}/D^2(m(x))$; $cv.v(res) = \sqrt{V(est(\sigma^2 I))}/\sigma^2 I$

Figure 7 shows as example, the sampling values $N_x=50$ and $N_e=50$ of one realisation for each of the four theoretical spatial distributions.

Table 3 provides the results for the ratios of the means and Table 4 those for the ratios of the standard deviations. When estimating the process mean M_L, there is virtually no effect of the allocation of sampling effort neither on the mean nor the CV. When estimating the process variance D^2_L, the situation is different. Allocating more sampling effort to sample finely the trend (N_x>=N_e) leads to a lower bias and a better precision. The residual variance is always over-estimated when the sampling effort is allocated predominantly at small scale (N_e>N_x). The variance of the estimate of the residual variance is also always higher for the sampling designs in which N_e>N_x. This is in coherence with the fact that the trend values play an important role in the equations of all variance terms.

3. Conclusion

When analysing coherence in the variance of the data sets coming from the IBTS and the box surveys, the question was raised whether the IBTS could under-estimate the overall (dispersion) variance and the variance within the ICES rectangles (i.e., small scale). The simulations undertaken using a model of spatial distribution with a trend highlighted the importance of the large scale sampling for estimating these variances. A survey design consisting in sampling intensively in a few locations would result in over-estimating the overall variance as well as the average variance at any given location (i.e., residual variance or small scale variance).

Table 1 : Basic statistics for North Sea cod age 2 quarter 2 in 1991

	IBTS	Average for all Boxes	Box.A	Box.B	Box.C	Box.D
nb.hauls	297	25.5	27	25	24	26
mean	5.55	9.18	8.89	9.83	3.78	17.42
variance	80.35	114.6	39.10	185.65	124.15	109.47
minimum	0	1.09	1.45	0	0	2.93
maximum	81.2	41.19	24.28	50.67	47.46	42.35
Percent zeroes	0.28	0.21	0	0.20	0.63	0

Table 2 : Variogram spherical model parameters for IBTS data. Nugget and Sill values are in squared numbers per hour (as variance parameters), Ranges are expressed in Nm.

Nugget	Sill.1	Range.1	Sill.2	Range.2
40	22	36	26	330

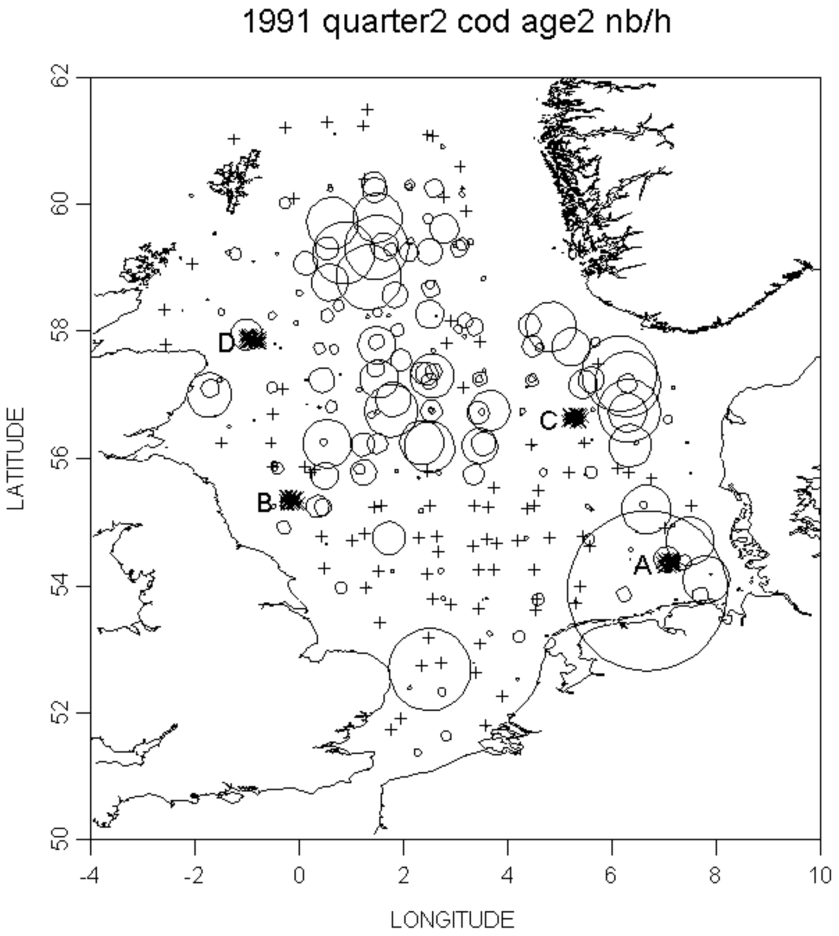

Fig. 1 : Proportional representation of the IBTS data. Crosses are zeroes.

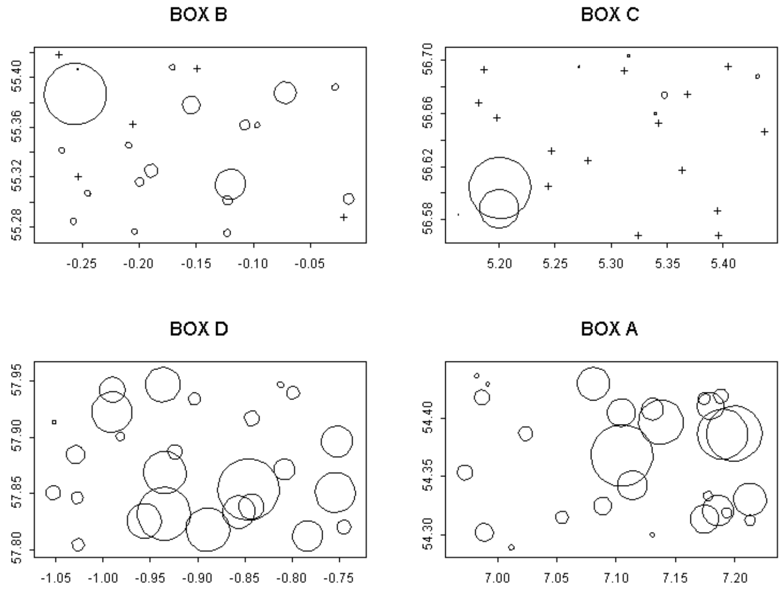

Fig. 2 : Proportional representation of the Boxes data. Crosses are zeroes. Circle radii are proportional to the maximum value of the pooled box data.

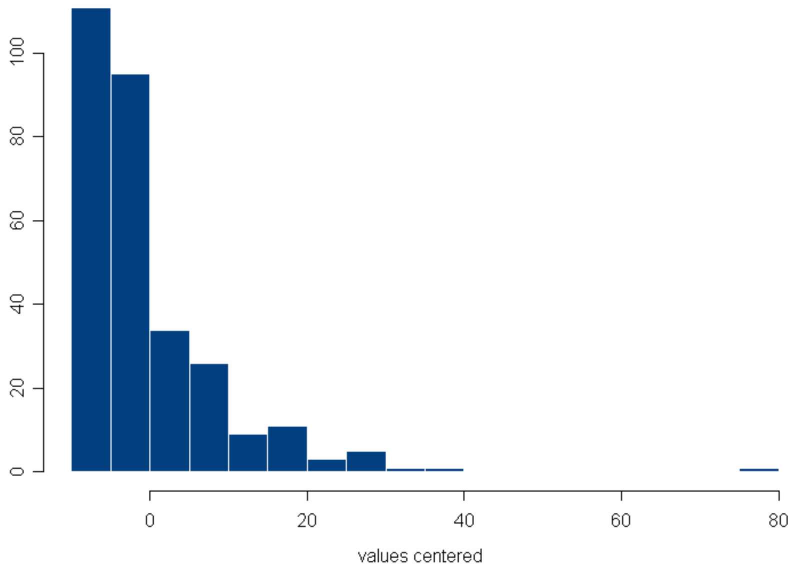

Fig. 3a : Histogram of the IBTS data. Values are centred around their mean.

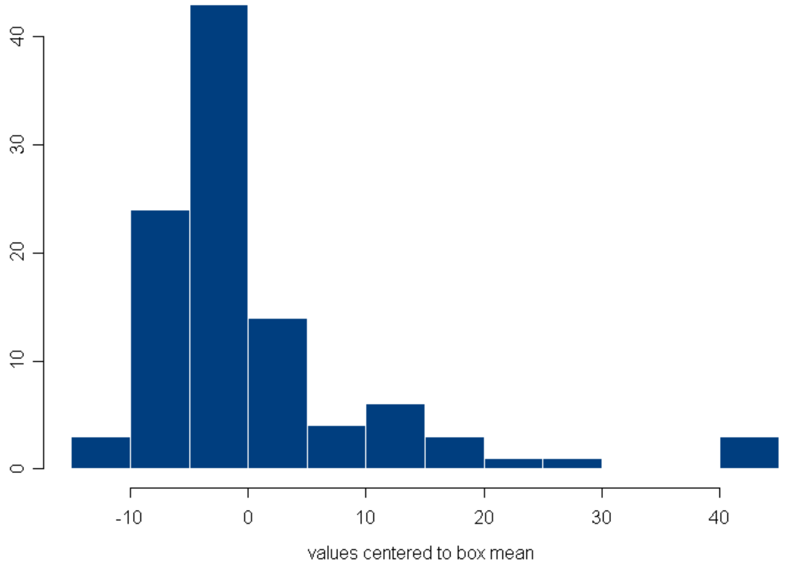

Fig. 3b : histogram of the pooled box data. Values for each box have been centered around the box mean.

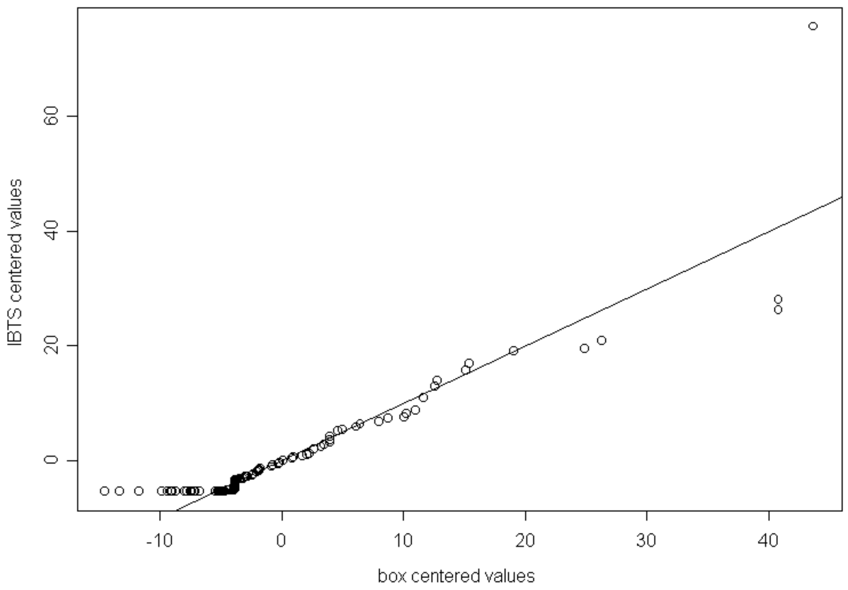

Fig. 4 : QQ-plot between histograms of IBTS and box data.

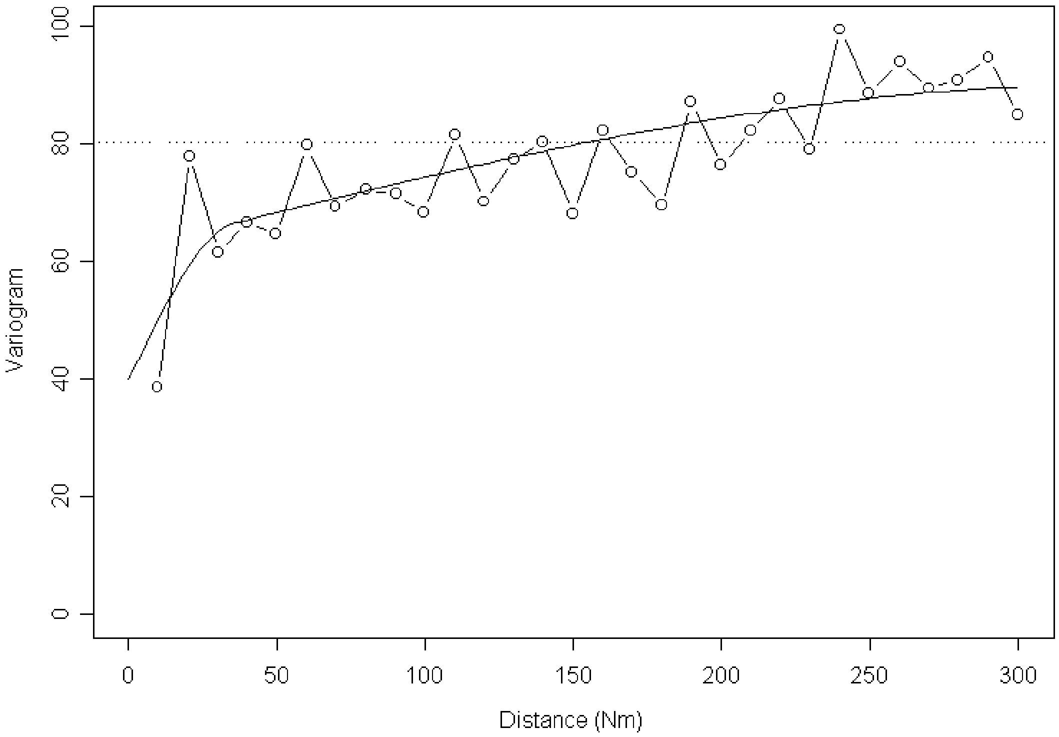

Fig. 5 : Variogram for the IBTS and its fitted model. Dashed line represents data variance.

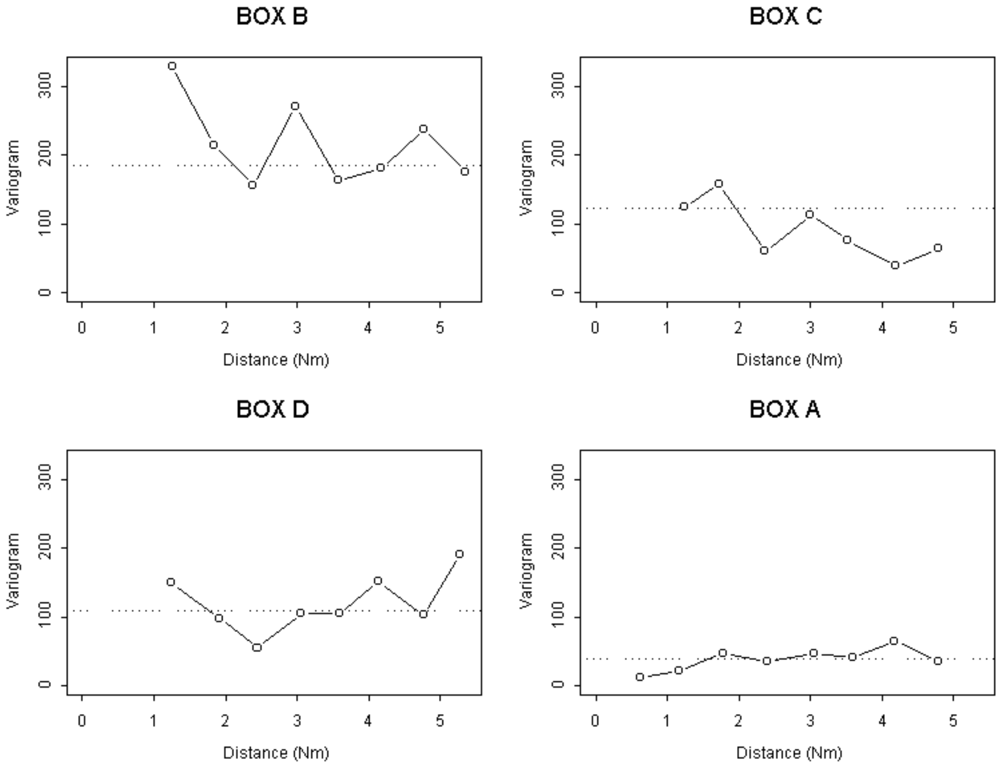

Fig. 6 : Variograms within the boxes for the Box data. Dashed lines represent data variance in each box.

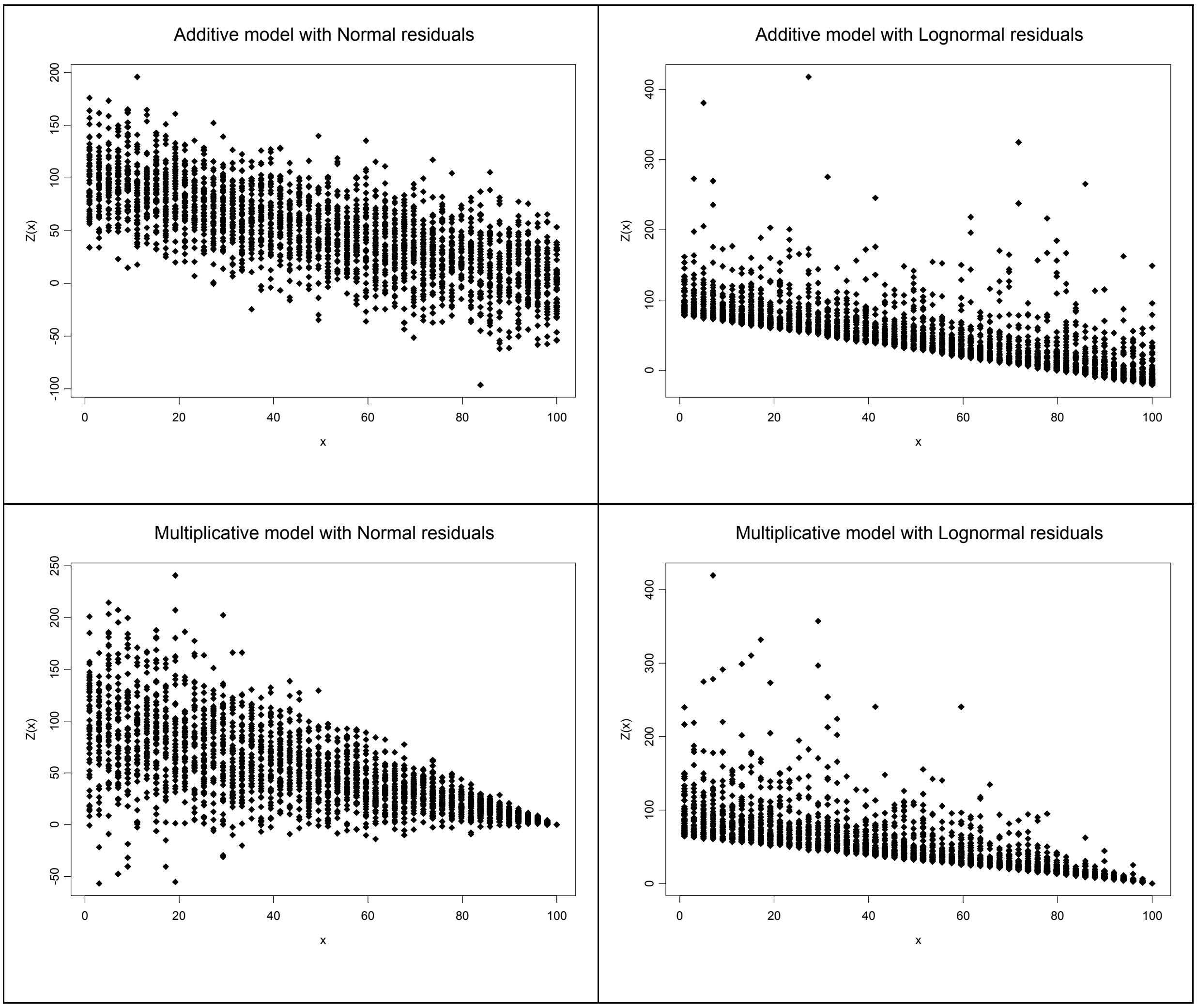

Fig. 7: Examples of simulated spatial distributions sampled with Nx=50 samples regularly spaced along x and Ne=50 replicated samples at each location.

Table 3: Estimation of process mean and variance terms for different models, sampling effort and allocation of the effort between large scale and small scale.
Nx: number of samples along the line
Ne: number of replicated samples at each sampling location Nx
$m.m(Z) = E(\hat{Z}_L)/M_L$; $m.v(Z) = E(\hat{V}_L)/D^2{}_L$
$m.v(trend) = E(\hat{D}^2(m(x)))/D^2(m(x))$; $m.v(res) = E(est(\sigma^2 I))/\sigma^2 I$

	m.m(Z)	m.v(Z)	m.v(trend)	m.v(res)	
Nx=14; Ne=14	0.98	1.08	1	1.3	Additive Model Normal Resiudals
Nx=50; Ne=4	0.99	1	1	1.26	
Nx=4; Ne=50	0.99	1.31	0.98	2.19	
Nx=7; Ne=7	0.99	1.16	1.03	1.63	
Nx=25; Ne=2	0.99	1.05	1	1.62	
Nx=2; Ne=25	0.99	2.02	1	5.98	
Nx=14; Ne=14	0.99	1.02	0.94	1.25	Additive Model Log-normal Resiudals
Nx=50; Ne=4	0.99	1	1.01	1.27	
Nx=4; Ne=50	0.98	1.31	0.96	2.23	
Nx=7; Ne=7	0.98	1.09	0.87	1.63	
Nx=25; Ne=2	0.98	1.02	0.93	1.6	
Nx=2; Ne=25	0.98	1.93	0.87	5.91	
Nx=14; Ne=14	1	1.08	1	1.31	Multiplicative Model Normal Resiudals
Nx=50; Ne=4	0.99	1.01	0.98	1.31	
Nx=4; Ne=50	0.99	1.4	1.15	2.2	
Nx=7; Ne=7	0.99	1.18	1.02	1.67	
Nx=25; Ne=2	0.99	1.05	1	1.62	
Nx=2; Ne=25	0.99	2.18	1.36	5.95	
Nx=14; Ne=14	0.99	1.04	0.97	1.26	Multiplicative Model Log-normal Resiudals
Nx=50; Ne=4	0.99	1	1	1.27	
Nx=4; Ne=50	0.98	1.3	0.98	2.16	
Nx=7; Ne=7	1	1.3	1.23	1.75	
Nx=25; Ne=2	0.99	1.12	1.17	1.68	
Nx=2; Ne=25	1	2.2	1.32	6.1	

Table 4: Standard deviation of process mean and variance estimates for different models, sampling effort and allocation of the effort between large scale and small scale.
Nx: number of samples along the line
Ne: number of replicated samples at each sampling location Nx

$cv.m(Z) = \sqrt{V(\hat{Z}_L)} / M_L$; $cv.v(Z) = \sqrt{V(\hat{V}_L)} / D^2{}_L$

$cv.v(trend) = \sqrt{V(\hat{D}^2(m(x)))} / D^2(m(x))$; $cv.v(res) = \sqrt{V(est(\sigma^2 I))} / \sigma^2 I$

	cv.m(Z)	cv.v(Z)	cv.v(trend)	cv.v(res)	
Nx=14; Ne=14	0.04	0.09	0.1	0.18	Additive Model Normal Resiudals
Nx=50; Ne=4	0.04	0.09	0.12	0.16	
Nx=4; Ne=50	0.04	0.1	0.1	0.25	
Nx=7; Ne=7	0.08	0.19	0.22	0.37	
Nx=25; Ne=2	0.08	0.18	0.28	0.31	
Nx=2; Ne=25	0.09	0.26	0.22	0.98	
Nx=14; Ne=14	0.04	0.28	0.55	0.14	Additive Model Log-normal Resiudals
Nx=50; Ne=4	0.04	0.35	0.74	0.2	
Nx=4; Ne=50	0.04	0.36	0.65	0.29	
Nx=7; Ne=7	0.07	0.4	0.77	0.38	
Nx=25; Ne=2	0.07	0.66	1.3	0.72	
Nx=2; Ne=25	0.07	0.43	0.8	0.97	
Nx=14; Ne=14	0.05	0.11	0.14	0.19	Multiplicative Model Normal Resiudals
Nx=50; Ne=4	0.04	0.11	0.15	0.17	
Nx=4; Ne=50	0.04	0.14	0.18	0.3	
Nx=7; Ne=7	0.09	0.23	0.35	0.45	
Nx=25; Ne=2	0.08	0.17	0.33	0.36	
Nx=2; Ne=25	0.1	0.38	0.47	1.22	
Nx=14; Ne=14	0.04	0.37	0.62	0.2	Multiplicative Model Log-normal Resiudals
Nx=50; Ne=4	0.04	0.39	0.7	0.29	
Nx=4; Ne=50	0.04	0.34	0.53	0.26	
Nx=7; Ne=7	0.1	1.43	2.59	0.76	
Nx=25; Ne=2	0.08	0.88	1.58	1	
Nx=2; Ne=25	0.1	0.88	1.19	1.28	