Point Clustering of Minke Whales in the Northeastern Atlantic

Gro Hagen
Norwegian Computing center, Box 114 Blindern, N-0314 Oslo, Norway

Tore Schweder
University of Oslo, Box 1095 Blindern, N-0317 Oslo

November 26, 1994

Abstract

A statistical model for spatial whale distribution in the Northeastern Atlantic is presented. The model, a Neyman-Scott cluster process, is fitted separately for 12 strata, based on data from the Norwegian shipborn survey in 1989.

The spatial distribution differs considerably between areas, both with respect to cluster size, intensity of clusters and density of whales within clusters. It seems obvious that information about the clustering is relevant when planning new surveys, and when analyzing survey data.

KEY WORDS: survey data, spatial distribution, cluster process

1 Introduction

The spatial distribution of minke whales is of interest for various reasons. The degree of clustering is of independent biological interest, both with respect to behaviour and feeding. The variability in the line transect surveys will depend on the degree of clustering, and the interpretation of survey data will be facilitated if a spatial point process model has been fitted.

We will use transect data from the Norwegian shipborn survey in 1989[6] to fit a Neyman-Scott cluster process, separately for each stratum. The estimation is done by fitting the theoretical K-function[1] to its empirical counterpart by numerical optimization. The K-function summarizes the interpoint distances in the point process.

2 Statistical model for whale distribution

The point process model for spatial whale distribution in the Northeastern Atlantic is a Neyman-Scott cluster model. When defining such a model, we need to specify a statistical model that covers both the cluster units and the individual whales belonging to each cluster. Separate models are fitted to each of the survey blocks used in Øien[6]. The model, and simulation, is based on:

Cluster units. The clusters form a stationary Poisson Process with intensity λ per unit area. This means that the number of clusters in a region A, with area $\nu(A)$ is Poisson distributed with parameter (and expected value) $m = \lambda \cdot \nu(A)$. The clusters' positions inside the region, represented by the positions of cluster centers, are randomly distributed over the total region according to a uniform probability distribution.

Individual whales belonging to a cluster. The individual whales of a given cluster is a Poisson Process, but with intensity decreasing as the distance from the cluster center increases. The Poisson intensity of whales in a cluster is

$$\mu \exp\left(-\frac{x^2+y^2}{2\rho^2}\right) \tag{1}$$

where x is the distance from the center in x-direction and y the distance in y-direction. The intensity model (1) specifies that the number of points in a cluster is Poisson distributed with parameter

$$M = \mu 2\pi\rho^2 \tag{2}$$

The points are independently scattered around the cluster center, according to the spherical bivariate normal distribution with variance in both x- and y-direction ρ^2.

The parameter ρ might be interpreted as the radius of the cluster, and the parameter μ is the whale intensity in the center. If ρ is small relative to the size of the area, the method of simulation indicated above works well. If, however, ρ is large, edge effects come into play.

Poisson Cluster Processes are described in Diggle[1] and Ripley[3]. The process consists of parent events (cluster centers), and to each parent there are a number of offspring (individual whales) with relative positions independently and identically distributed according to a probability distribution (bivariate normal). The intensity of parent events is λ, and the number of offspring for each parent is a Poisson variable with parameter M, given in (2).
To summarize the interpoint distances in a Poisson Cluster Process, one introduces the so-called K-function, defined by

$$\gamma K(t) = \text{expected number of further points within distance } t \text{ from an arbitrary point in the process}$$

where γ is the overall intensity of points,

$$\gamma = \lambda M = \lambda\mu 2\pi\rho^2 \tag{3}$$

In the present model, "points" are the individual whales, and $\gamma K(t)$ thus is defined as the expected number of further whales within an interpoint distance not exceeding t from a randomly chosen whale. From the theory in Diggle[1] we find that in the present model,

$$K(t) = \pi t^2 + \frac{1}{\lambda}(1 - \exp(-\frac{t^2}{4\rho^2})) \tag{4}$$

and so the μ-parameter disappears.

3 Estimation

To estimate the model parameters we will use the K-function, and we therefore need an empirical point estimate for $K(t)$ based on observed data. The estimator $\widehat{K}(t)$ is described in Diggle[1] and Ripley[3], and we only include a short summary. The function $\gamma K(t)$ is the expected number of points (whales) within a distance t from a randomly chosen point (whale). For a general region A with area $\nu(A)$, the expected number of points is $\gamma\nu(A)$. The expected number of ordered pairs of points, with the first point being inside A thus become

$$N_P(t) = \gamma\nu(A) \cdot \gamma K(t) = \gamma^2\nu(A)K(t) \tag{5}$$

The method for estimating $K(t)$ is to estimate $N_P(t)$, and then to combine (5) with the common estimator for overall intensity $\widehat{\gamma} = \frac{n}{\nu(A)}$, to get an estimator for the K-function:

$$\widehat{K}(t) = \frac{\nu(A)}{n^2}\widehat{N}_P(t)$$

where n is the total number of points observed inside the region A. Let $P_1, \ldots, P_I$ be the process points inside the region A, and let u_{ij} be the distance between P_i and P_j. Following Ripley[3], we estimate N_P as a weighted sum of ordered pairs, with weights being the inverse of the probability that a point at distance u_{ij} from the point P_i will be observed (i.e. is inside the region A). When the process is stationary and isotropic, as is the case in our model, this probability is the proportion of the circumference with radius u_{ij} and center at P_i that lies inside the observed region A.

Our observed points are from line transects through relatively large areas. It seems realistic to define the observed region (region A in the text above) to be a small strip of width b around the transect line. The total observed area in a block (stratum) is

$$\nu(A) \approx hTb$$

where h = intended vessel speed during transect, T = transect duration inside the actual block and b the effective search width on the transect. In 1989 the intended vessel speed was $h = 10$ knots, and the effective search width has been estimated to approximately 0.18 nautical miles[4]. When estimating the model parameters, we therefore use these values in all formulas.

In several of the blocks the data consists of observations from more than one continous transect. We have chosen to regard these as independent observation series from the same model. They are therefore linked together in one continous transect, covering the actual block. The only relevant parameters are block identification and relative time of observation. From these we can compute a theoretical position of each observation, and perform the estimation. For one specific block, let the relative points of time for whale observations be $0 \leq \tau_1 \leq \tau_2 \leq \ldots \leq \tau_n < T$. Relative forward positions in the observed strip will then be $0 \leq h\tau_1 \leq h\tau_2 \leq \ldots \leq h\tau_n < hT$, and these can also be regarded as the observed absolute positions along the transect strip. The reason for this is that the effective search width is very small compared to the forward distances between observations. The K-function and its estimator $\widehat{K}(t)$ thus become

$$\begin{aligned} K(t) &= K(h\tau) = \pi(h\tau)^2 + \frac{1}{\lambda}\left(1 - \exp(-\frac{(h\tau)^2}{4\rho^2})\right) \\ \widehat{K}(t) &= \widehat{K}(h\tau) = \frac{h^2\pi T}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mid \tau_i - \tau_j \mid I_t(u_{ij}) \end{aligned} \tag{6}$$

where $I_t(u) = 1$ if $u < t$, and zero else. To stabilize the variance, Ripley[3] suggests to use $L(t) = \sqrt{\frac{K(t)}{\pi}}$ when estimation is done by least squares. For a given survey block, the estimates of λ and ρ are those values which minimize the squared error:

$$\sum_{t\leq\Delta t}\left(L(t) - \widehat{L}(t)\right)^2$$

However, the above functions do not include the μ-parameter. Since the expected number of observed points along the transect is $E(n) = \lambda\mu 2\pi\rho^2 \cdot hTb$, all the parameters are estimated by minimizing

$$Q(\Delta t) = \sum_{t\leq\Delta t}\left(L(t) - \widehat{L}(t)\right)^2 + W\left(\lambda\mu 2\pi\rho^2 hTb - n\right)^2 \tag{7}$$

Here W is a constant, weighting the L-fitting and the fitting of the expected total to the observed total, n. If W is chosen big, the result is approximately equivalent to minimizing the sum under the constraints that $\lambda\mu 2\pi\rho^2 hTb = n$. During the estimation we have chosen this kind of weighting, by setting $W = 100$. The function (7), and consequently the estimates, depend on the window size Δt. Our main interest is in the rather local clustering of whales, and so we will concentrate on observations made close in time (and distance). We have found that a Δt-value of approximately 10 nautical miles (that means a maximum time difference of 60 minutes) is close to optimal, in the sense that this gives relatively good fit for the majority of the survey blocks. We want to use the same Δt when estimating the model in all blocks, and

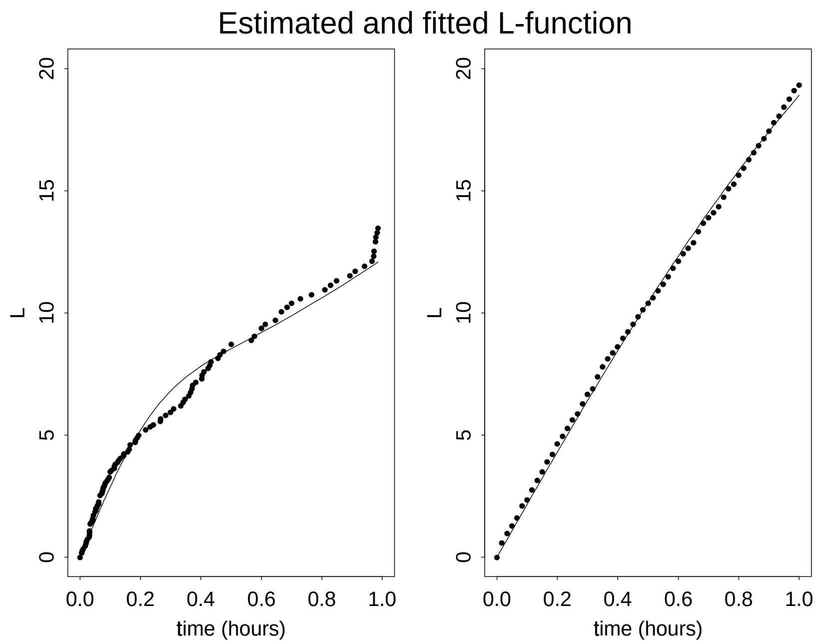


Figure 1: Empirical L-function (dots) and fitted L-function (smooth curve), from the blocks VSS (West of Spitsbergen, south) and KO (Kola).

therefore all estimates presented are those minimizing the function (7) when $\Delta t = 10$ nautical miles. The data points included in the sum (6), are those present in the actual observation series. The estimation results (blockwise parameter estimates) are presented in table 1. The parameter estimates differ a lot between blocks. Figure 2 shows two regions of 10000 square nautical miles area, simulated according to the two estimated models from the blocks VSS (West of Spitsbergen, south) and KO (Kola).

4 Discussion

The model parameters (μ, ρ, λ) were estimated by numerical minimization of (7), separately for each of the 12 blocks. The parameter estimates are presented in table 1, and from this table we observe that the parameter estimates differ a lot. Figure 1 shows that the L-functions are well fitted to the empirical estimates, and it also illustrates the dilemma when deciding the actual window size Δt. In the upper tail of the left figure there is a substantial increase in the empirical values, which is not reflected in the fitted function. This sudden increase can be interpreted as an inclusion of points from different clusters in the sum defining the empirical L-function, and is therefore of minor interest when investigating the individual clusters. The different sets of parameter estimates give rise to varying spatial whale distributions. The model estimated for block VSS (West of Spitsbergen, south) is an example of a model with many small clusters of high whale intensity, while the KO-model (Kola block) is an example of a model with relatively few and large clusters in terms of radius, and of high whale intensity. Block SV (Svalbard area, southwest of the Spitsbergen blocks) has parameter estimates very different from the other blocks. The fitted model for this block has very few clusters of enormous radius but with very small whale density.

Block	$\widehat{\mu}$	$\widehat{\rho}$	$\widehat{\lambda}$
VSN	3.131	1.230	$8.896 \cdot 10^{-3}$
VSS	5.951	1.244	$6.539 \cdot 10^{-3}$
SV	0.292	133.979	$5.074 \cdot 10^{-6}$
BJ	1.630	2.161	$2.569 \cdot 10^{-3}$
BA	9.765	5.815	$1.181 \cdot 10^{-4}$
GÅ	2.351	2.617	$9.951 \cdot 10^{-4}$
KO	6.446	5.720	$6.601 \cdot 10^{-4}$
FI	1.363	1.886	$5.327 \cdot 10^{-3}$
LO	1.642	2.156	$4.425 \cdot 10^{-3}$
NØ	2.024	2.861	$6.966 \cdot 10^{-4}$
SN	0.932	2.667	$2.479 \cdot 10^{-3}$
NS	0.762	16.684	$1.049 \cdot 10^{-4}$
Median value	1.833	2.642	$1.737 \cdot 10^{-3}$

Table 1: Blockwise Parameter estimates based on observations from the Norwegian survey in 1989, see Øien[6] for block identification.

When the models are to be interpreted, one must have in mind that the estimated whale distribution is a theoretical model for regions of unlimited area. The actual blocks then must be regarded as a smaller region, randomly placed in the infinite area of the theoretical model. However, the result is clearly that the whale distribution in the Northeastern Atlantic can be modelled as a Neyman-Scott process, with substantial clustering. Information on degree of clustering within the various survey blocks will be of help when designing future surveys. It is important to optimize the design so as to obtain abundance estimates of minimal variance, given total costs. Since variability depends on the degree of clustering, such information is helpful. From data gathered in a new survey, the Neyman-Scott model should be estimated for the individual survey blocks. This fitted models will be useful when the survey process is to be simulated, in order to estimate hazard probability paramters and effective search width, see[5].

References

[1] Diggle PJ. Statistical Analysis of Spatial Point Patterns. London: Academic Press, 1983

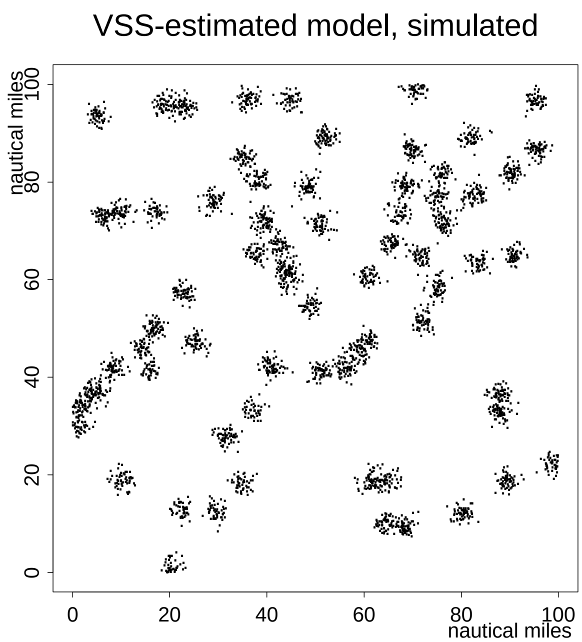
[2] Neyman J, Scott EL. A theory of spatial distribution of galaxies. Astrophysical J. 1952; 116: 144-163.

[3] Ripley BD. Spatial Statistics. New York: Wiley, 1981

[4] Schweder T, Øien N, Høst G. Estimates of Abundance of Northeastern Atlantic Minke Whales in 1989. Rep. int. Whal. Commn. 1993; 43: 323-331

[5] Schweder T, Hagen G. Effective search width in shipborn surveys of minke whales in the northeastern Atlantic: Concepts and methods. Contribution to the International Symposium on the Biology of Marine Mammals in the Northeast Atlantic, in Tromsø 29 November – 1 December 1994.

[6] Øien N. Abundance of the northeastern Atlantic stock of minke whales based on shipboard surveys cionducted in July 1989. Rep. int. Whal. Commn. 1991; 41: 499-511



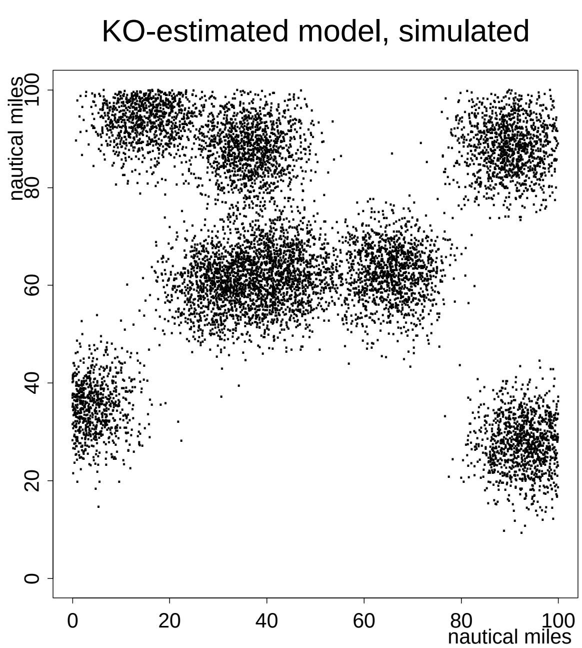


Figure 2: Two regions of 10000 sqnml simulated according to the estimated model from block VSS (West of Spitsbergen, south) and KO (Kola), respectively.