Estimation of a Neyman-Scott process from line transect data by a one-dimensional K-function

Magne Aldrin

Norwegian Computing Center,

P.O. Box 114 Blindern, N-0314 Oslo, Norway

email: magne.aldrin@nr.no

Marit Holden

Norwegian Computing Center (as above)

email: marit.holden@nr.no

Tore Schweder

Norwegian Computing Center (as above) and

Department of Economics, University of Oslo,

P.O. Box 1095 Blindern, N-0317 Oslo, Norway

email: tore.schweder@econ.uio.no

Summary

We consider the problem of estimating the parameters of a two-dimensional Neyman-Scott process, from data collected through a line transect survey. Cowling (1998) suggested an estimation method based on a one-dimensional K-function along the transect line. However, her expression for the theoretical K-function was wrong. We develop the correct K-function. We further carry out a simulation study, where we show that the one-dimensional K-function method outperforms a method previously suggested by Hagen and Schweder (1995).

Keywords: K-function, Line transect survey, Neyman-Scott process

1 Introduction

The spatial (here two-dimensional) distribution of a biological population may often be modelled by a spatial point process. We will consider one such process, the Neyman-Scott process, which is suitable for modelling clustered populations. Points (or objects) are observed by a line transect survey. The probability of detecting points decreases rapidly as a function of the perpendicular distance x from the transect line, so only points close to the transect are detected. Our aim is to estimate the parameters in the Neyman-Scott process from such data.

One way of characterising the spatial dependency in a point process is by the so called K-function. If all points are observed within a region, the model parameters may be estimated by fitting the theoretical K-function to its empirical counterpart (Ripley, 1977;. Diggle, 1983 and Cressie, 1991). However, when the points are observed by a line transect survey, the underlying Neyman-Scott process is thinned to a nearly one-dimensional point process.

Hagen and Schweder (1995) proposed an estimation method for use in line transect surveys for minke whales. They constructed a rectangle centred along the transect line, with the same length as the transect, and width equal to the area under the detection curve (the effective strip width), and assumed perfect observation within the rectangle. Then they projected the detected points onto the transect line (the y-coordinates). Finally, they estimated the parameters by fitting a two-dimensional theoretical K-function to a two-dimensional empirical K-function.

Thus, even if they used only one dimension of the data, they used a two-dimensional K-function.

The rationale for ignoring the x-coordinates is that the distance to the transect line is usually considerable shorter than the transect length, and that most of the information about the spatial pattern is contained in the coordinate along the transect. Cowling (1998) attempted to construct a wholehearted one-dimensional estimation method, based on a real one-dimensional K-function along the transect line. However, there are a couple of serious errors in the theoretical development of the method, leading to a wrong K-function. Cowling (1998) tested the method in a simulation study, but did not compare the results with the method of Hagen and Schweder (1995).

In this paper, we follow the idea of Cowling (1998), but develop the correct K-function. Then we carry out the same simulation study as Cowling (1998), with some small modifications, where we compare the corrected version of Cowling's method to the method of Hagen and Schweder (1995). Our corrected version of Cowling's method clearly outperforms the method of Hagen and Schweder (1995).

2 Parameter estimation of the Neyman-Scott process

2.1 The Neyman-Scott process and the detection function

The definition of a Neyman-Scott process is found in for instance Cressie (1993, p. 662). We will consider the following special version:

- Invisible parent events are Poisson distributed with intensity λ (per unit area).
- Each parent independently produces a Poisson(μ) number of offspring.
- The positions of the offspring relative to their parents are independent and have an isotropic bivariate normal distribution with variance ρ^2 in both x- and y-direction.

The detection function g(x) is the probability of detecting an offspring at a distance x from the transect line. We will assume a normal detection function,

$$g(x) = g_0 \cdot \exp(-x^2/2\sigma^2) \ , \tag{1}$$

where $g_0 = g(0)$ is the detection probability at x=0. The two parameters g_0 and σ are typically estimated from external data (Buckland et. al. 1993, Schweder et. al. 1999), and are assumed known in the present context. Cowling (1998) assumes $g_0 = 1$, but this assumption is unnecessary. For North-Atlantic minke whales g_0 is around 0.35 due to diving etc. The effective strip width is

$$\int_{-\infty}^{\infty} g(x) = \sqrt{2\pi}\sigma g_0 = 2\omega \ , \tag{2}$$

where ω is the effective strip half-width.

Suppose that a line transect survey of infinite length is carried out along x=0, and consider a parent located at x=c. Let T denote the detected offspring in that clus-

ter. Cowling (1998) shows that conditioned on c, the expected number of detected offspring is

$$E(T|c) = g_0 \mu p_c \ , \tag{3}$$

where $p_c = \sigma(\sigma^2 + \rho^2)^{-1/2} \exp\{-c^2/2(\sigma^2 + \rho^2)\}$. Further, the conditional distribution of T is Poisson, i. e. $T|c \sim Poisson(g_0 \mu p_c)$.

2.2 Parameter estimation

The plan is now to project the detected points onto the transect line, calculate the theoretical K-function for the resulting one-dimensional process, and fit it to the empirical one-dimensional K-function.

The K-function (Ripley, 1977) of a stationary spatial point process with intensity τ is defined as

$$K(h) = \tau^{-1} E\begin{pmatrix}\text{number of extra events within distance h} \\ \text{of a randomly chosen event}\end{pmatrix} . \tag{4}$$

Cressie (1993, p.665) gives the K-function for a d-dimensional Neyman-Scott process. In one dimension this becomes

$$K_1(h) = 2h + \frac{E\{T(T-1)\}}{\lambda_1 \{E(T)\}^2} F(h) \ , \tag{5}$$

where F(h) is the distribution function of the distance between two points in the same cluster, λ_1 is the intensity of the one-dimensional parent process, and T is the number of offspring in a cluster.

The one-dimensional process will still fulfil the conditions for a (stationary and isotropic) Neyman-Scott process. The offspring will be normally distributed around the cluster centres, and

$$F(h) = 2\Phi(h/\rho\sqrt{2}) - 1 \ , \tag{6}$$

where Φ is the distribution function of the standard normal distribution (Cowling 1998).

However, even if T is Poisson conditional on the x-coordinate of the cluster (x=c), the marginal distribution of T is *not* Poisson. Cowling (1998) wrongly assumes this. Her expressions for $E\{T(T-1)\}/(\lambda_1\{E(T)\}^2)$ and K_1 are thus wrong. The correct K-function for the detected points projected onto the transect line is

$$K_1(h) = 2h + \{2\Phi(h/\rho\sqrt{2}) - 1\}/(2\sqrt{\pi}\lambda\sqrt{\sigma^2 + \rho^2}) \ , \tag{7}$$

see Appendix A for a proof.

In Figure 1 we show Cowling's K_1-function and the corrected K_1-function for three set of parameter values, for h up to $h_0 = 5\rho$. The relation between σ and ω is given by (2), with $g_0 = 1$ in all three sets of parameter values.

Insert Figure 1 about here.

The empirical K-function is estimated by

$$\hat{K}_1(h) = 2n^{-2}L\sum_j\sum_{i<j} I(|y_i - y_j| < h) \ , \tag{8}$$

where L is the length of the transect, n is the number of detected points, and y_i and y_j are the positions along the transect line. Note that since the theoretical K-function is developed for an infinitely long transect, there will be a systematic end effect in the difference between $K_1(h)$ and $\hat{K}_1(h)$. We will not try to correct for this.

Now, the parameters λ and ρ can be estimated by minimising

$$\int_0^{h_0} [\sqrt{\hat{K}_1(h)} - \sqrt{K_1(h;\lambda,\rho)}]^2 dh \ , \tag{9}$$

for a suitable value of h_0. The square root transformation was suggested by Diggle (1983) as reasonable when the point process is close to Poisson, but other transformations could be more effective for more clustered processes.

We further have that

$$\mu = ((E(n))/(\sqrt{2\pi}\lambda\sigma g_0 L)) \ , \tag{10}$$

(see Appendix A for proof), such that μ may be estimated by substituting E(n) by the observed n,and λ and σ by their estimates.

We have so far assumed that σ is known. However, from (7) we see that $K_1(h)$ depends on σ in addition to λ and ρ. Therefore, it should be possible to estimate all these parameters from (9). However, in certain situations σ may be better estimated by other data sources, as described in Schweder et. al. (1999).

3 Simulation study

Cowling (1998) conducted a simulation study to investigate the bias and variance of the parameter estimates of her method. We have repeated this simulation study with some small modifications.

For each of 30 different sets of parameter values we carried out simulation experiments numbered from 1 to 30. The parameter values are given in Table 1, except for g_0, which was 1 in all experiments. For each simulation experiment, the length L of the transect line was chosen such that the expected number of detected points was 250, i.e. $L = 250/\sqrt{2\pi}\lambda\mu\sigma$. Then 250 samples (replications) of observations were simulated by the following procedure: First, cluster centres were simulated in a rectangle centred along the transect line, with length L and width 2b, where the width 2b was very large compared to ρ and σ. Then the points were simulated, but only points inside the rectangle were kept. Finally, the points were detected randomly according to the detection curve (1). This procedure has one disadvantage: Only points belonging to cluster centres within the rectangle will be simulated. This means that we potentially loose some points belonging to cluster

centres just before the start or after end of the transect line. The procedure above may differ slightly from that of Cowling (1998), since Cowling does not explain her procedure in detail.

Insert Table 1 about here.

Only samples with at least 60 detected points were used for estimation. From the remaining samples the parameters λ, μ and ρ were estimated by the method of Hagen and Schweder (1995) and our corrected K_1, whereas σ was assumed to be known. As Cowling (1998), we employed the limit of integration $h_0 = 5\rho$, though in practice h_0 has to be chosen without knowing the true parameters. The parameter estimates of each method were found by numerical optimization of (9). Unfortunately, the results depend on the starting values of the parameters. We therefore started the optimization from 12 different sets of starting values, and used the most optimal solution as the final parameter estimates. The 12 starting values were $\lambda = 0.001, 0.003, 0.009, 0.027$ combined in all possible ways with $\rho = 1.0, 5.5, 10.0$. The corresponding value of μ is given by (10) with E(n) replaced by the observed n. Cowling (1998), on the other hand, used only the true parameter values as starting values. This could give over-optimistic results if the criterion function in (9) is unsmooth, and may favour methods that have this undesirable property that the optimization criterion is unsmooth.

Sometimes, the estimates could be very small or very large for one or more of the methods. Samples with parameter estimates outside the range $(10^{-20}, 10^{20})$ for at least one of the methods were excluded from the comparisons between methods.

For the method of Hagen and Schweder (1995), one or more samples had parameter estimates outside this range in 8 of the 30 simulation experiments, whereas this happened only in 1 experiment for the corrected K_1-function.

The parameter estimates are compared to the nominal parameter values. Note, however, that since we use only samples with at least 60 observations, and with parameter estimates within the range $(10^{-20}, 10^{20})$, the true parameter values may in fact differ from the nominal ones. We will however neglect this fact.

Cowling (1998) presented bias and variance for her method, but as a single measure we prefer to use RMSE = root mean squared error = $\sqrt{\text{variance} + \text{bias}^2}$. For each simulation experiment we have calculated RMSE for each method and for each parameter, and compared the two methods through the log ratio $\log(\text{RMSE}^{h\&s}/\text{RMSE}^{corr})$. Here “h&s” denotes the method of Hagen and Schweder (1995) and “corr” the corrected K_1-method. A positive value of this measure means that the the former method is worse than the latter, and vice versa. The results are shown in Figure 2Figure 2. The method with the corrected K_1-function is best in almost all situations, except for μ and ρ in experiment number one.

Insert Figure 2 about here.

4 Conclusions

We have followed the ideas of Cowling (1998) and corrected her expression for the one-dimensional K_1-function. The simulation study showed that this version with the corrected K_1-function did improve on the former method of Hagen and Schweder (1995). This was one of the intentions of Cowling's work.

Cowling (1998) also refers to another estimation method by Brown and Cowling (1998), and claims that their method is better than the K-function method of Cowling (1998). However, Brown and Cowling (1998) compare the two methods in 6 simulation experiments with different parameter values. If we calculate $\log(\text{RMSE}^{\text{Brown\&Cowling}}/\text{RMSE}^{\text{Cowling}})$ we get (-1.7, -0.4, 0.2, -0.4, -0.1, -0.2) for λ, (0.6, 0.7, 1.1, -0.3, -0.1, 0.0) for μ and (0.4, 0.9, 1.3, 0.0, 0.2, 0.5) for ρ. From these results, one can not conclude that the method of Brown and Cowling (1998) is better than Cowling's K-function method. Though, it should be noted that in the method of Brown and Cowling (1998), σ is estimated as well, whereas it is (probably) assumed known in the K-function method of Cowling (1998).

Acknowledgements

This paper has been supported by The Research Council of Norway, grant no. 121144/420. We thank David Hirst for helpful suggestions on drafts of this article.

Appendix A Proof of eq. (7) and (10)

To prove equation (7) we will need the following results

$$\int_{-\infty}^{\infty} p_c dc = \sigma\sqrt{2\pi} \ , \tag{11}$$

$$\int_{-\infty}^{\infty} p_c^2 dc = \sigma^2\sqrt{\pi}/\sqrt{\sigma^2+\rho^2} \ , \tag{12}$$

which are easily shown by using that the normal distribution density integrates to 1. Further, if $Z \sim Po(m)$,

$$E\{Z(Z-1)\} = m^2 \ . \tag{13}$$

Assume now first that we only consider clusters with centres at x=c within the range (-C,C). Then the intensity of the one-dimensional parent process is $\lambda_1 = 2C\lambda$.

All c-values within (-C,C) are equally probable, so when we take the expectation over c, c is treated as uniform with density 1/(2C) between -C and C. We then get

$$E(T) = E[E(T|c)] = E(g_0\mu p_c) = g_0\mu E(p_c) = g_0\mu \int_{-C}^{C} \frac{1}{2C} p_c dc \ , \tag{14}$$

and from (13)

$$E\{T(T-1)\} = E[E\{T(T-1)|c\}] = E(g_0^2\mu^2 p_c^2) = g_0^2\mu^2 E(p_c^2) = g_0^2\mu^2 \int_{-C}^{C} \frac{1}{2C} p_c^2 dc \quad (15)$$

This gives

$$\frac{E\{T(T-1)\}}{\lambda_1\{E(T)\}^2} = \frac{g_0^2\mu^2(2C)^{-1}\int_{-C}^{C} p_c^2 dc}{2C\lambda g_0^2\mu^2(2C)^{-2}\left(\int_{-C}^{C} p_c dc\right)^2} = \frac{\int_{-C}^{C} p_c^2 dc}{\lambda\left(\int_{-C}^{C} p_c dc\right)^2} \quad . \quad (16)$$

Letting $C \to \infty$, (7) is then obtained from (5), (6), (11) and (12).

To prove (10), assume again that we only consider clusters within (-C,C). Then the expected number of detected points on a transect line of length L are

$$E(n) = L2C\lambda E(T) = L2C\lambda g_0\mu \int_{-C}^{C} \frac{1}{2C} p_c dc \quad . \quad (17)$$

Letting C go to infinity and solving for μ gives (10).

References

Buckland, S.T., Anderson, D. R., Burnbaum K. P. and Laake, J. L. (1993). *Distance Sampling: Estimating Abundance of Biological populations*. London: Chapman and Hall.

Brown, B. M. and Cowling, A. (1998). Clustering and abundance estimation for Neyman-Scott models and line transect surveys. *Biometrika* **85**, 427-438.

Cowling, A. (1998). Spatial Methods for Line Transect Surveys. *Biometrics* **54**, 828-839.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised Edition)*. New York: Wiley.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Hagen, G. and Schweder, T. (1995). Point clustering of minke whales in the north-eastern Atlantic. In *Whales, Seals, Fish and Man*, A. S. Blix, L. Walløe and Ø. Ulltang (eds), 27-33. Amsterdam: Elsevier Science B.V.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 172-212.

Schweder, T., Skaug, H. J., Langaas, M. and Dimakos, X. K. (1999). Simulated Likelihood Methods for Complex Double-Platform Line Transect Surveys. *Biometrics* **55**, 678-687

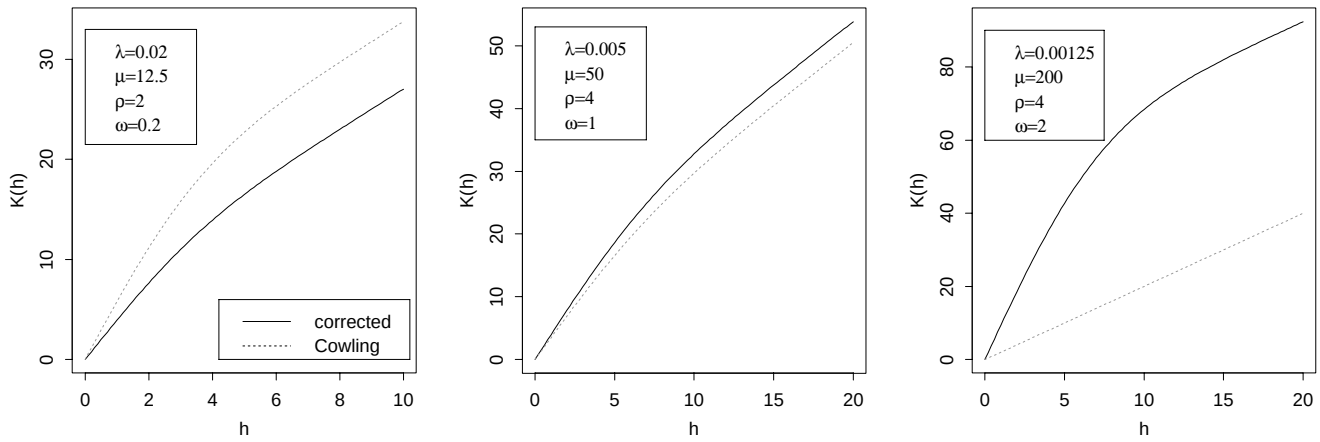

Figure 1 Plot of Cowling's K_1-function and the corrected K_1-function for various parameter values.

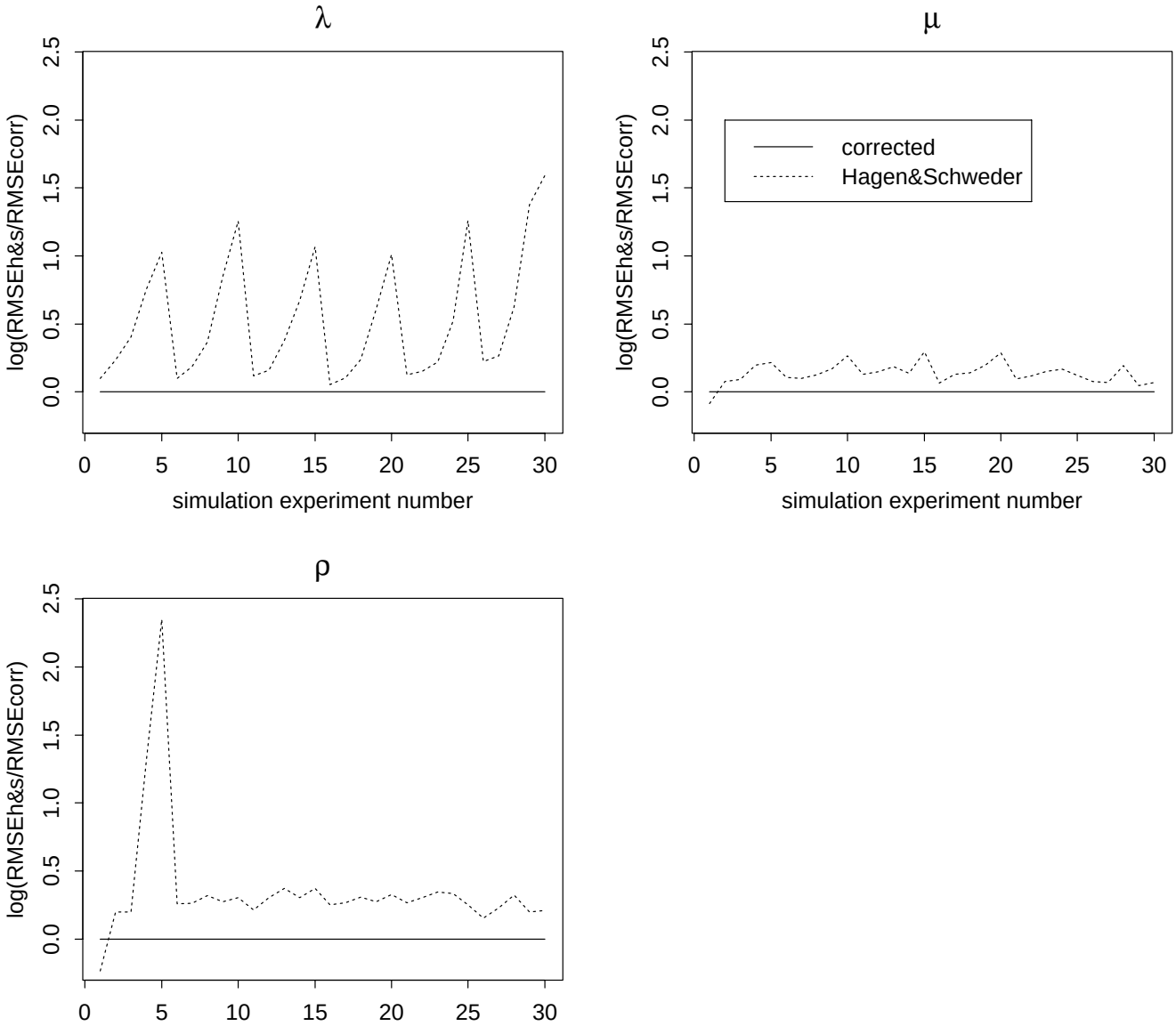

Figure 2 Log($RMSE^{h\&s}/RMSE^{corr}$).

Table 1 Parameter values and numbering of simulation experiments.

λ	μ	ρ	ω=0.2	ω=0.5	ω=1.0	ω=2.0	ω=3.0
0.02000	12.5	2	1	2	3	4	5
0.00500	50	2	6	7	8	9	10
0.00125	200	2	11	12	13	14	15
0.00500	50	4	16	17	18	19	20
0.00125	200	4	21	22	23	24	25
0.00125	200	8	26	27	28	29	30