Estimating the relationship between abundance and distribution

Rindorf, Anna; Lewy, Peter

Published in:
Canadian Journal of Fisheries and Aquatic Sciences

Link to article, DOI:
10.1139/F2011-153

Publication date:
2012

Document Version
Peer reviewed version

Citation (APA):
Estimating the relationship between abundance and distribution

Anna Rindorf and Peter Lewy

Abstract: Numerous studies investigate the relationship between abundance and distribution using indices reflecting one of the three aspects of distribution: proportion of area occupied, aggregation, and geographical range. Using simulations and analytical derivations, we examine whether these indices provide unbiased estimates of the relationship when estimated from count data. The indices investigated include the proportion of empty samples, the proportion of structurally empty samples, Lloyds index of patchiness, measures derived from Lorenz curves (such as $D_{95}$ and the Gini index), and measures based on Euclidean distance to the centre of gravity of the spatial distribution. Only the proportion of structurally empty areas, Lloyds index, and indices of the distance to the centre of gravity of the spatial distribution are unbiased at all levels of abundance. The remaining indices generate relationships between abundance and distribution even in cases where no underlying relationships exists, although the problem decreases for measures derived from Lorenz curves when samples contain more than four individuals on average. To illustrate the problem, the indices are applied to juvenile North Sea cod, Gadus morhua.

Introduction

The relationship between abundance and distribution is a key issue in fisheries management and ecology, as a population that contracts into a small area at low abundance becomes highly vulnerable to both exploitation and climatic changes. The contraction renders catch rates high in spite of stock declines and has facilitated the overexploitation of several fish stocks (Hamre 1978; Beverton 1990; Hutchings 1996). Furthermore, the presence or absence of contraction affects the success of different management measures. A contracting stock can be safeguarded through protecting the area where the stock is concentrated, whereas stocks that become increasingly dispersed as density declines will not obtain the same beneficial effect of area-based management. In contrast, measures such as effort limitation may be efficient for non-contracting stocks but have little effect on contracting stocks, as sustained high catch rates will cause fishing mortality per unit effort to increase as abundance declines. The relationship is of even greater importance to stocks faced with climatically induced decreases in stock size; the decrease in abundance increases the risk of local extinction and the decrease in the number of local aggregations increases the risk of stock extinction (“double jeopardy”; Lawton 1996). Hence, detecting the presence or absence of a concentration of the stock as abundance declines is of great importance to successful fisheries management in a changing environment.

Numerous studies have attempted to determine whether a relationship between abundance and distribution exists in natural populations. Most of these studies are based on measuring one or several of three different aspects of distribution: the proportion of the distribution area occupied, aggregation, and geographical range. In the case where the individuals can be detected and counted without error, these aspects can be estimated directly and related to abundance. However, in the majority of marine cases, complete census is impossible. Instead, distribution is inferred from samples collected by nets, trawls, and traps or from counts of individuals in a specified area of the bottom. In these cases, the number of individuals detected varies as a result of both random variation associated with sampling and as a result of differences in local abundance.

Received 5 April 2011. Accepted 19 October 2011. Published at www.nrcresearchpress.com/cjfas on 7 February 2012. J2011-0106

Paper handled by Associate Editor Kenneth A. Rose.

A. Rindorf and P. Lewy, Technical University of Denmark, National Institute of Aquatic Resources, Charlottenlund Castle, 2920 Charlottenlund, Denmark.

Corresponding author: Anna Rindorf (e-mail: ar@aqua.dtu.dk).
Spatial statistical analyses can be used to separate sampling variation from changes in local abundance (Rivoirard et al. 2000; Stelzenmüller et al. 2005; Lewy and Kristensen 2009). However, to demonstrate a relationship between abundance and distribution, simpler distribution indices are frequently used. It is essential for the interpretation of the results based on such indices that random variation due to sampling does not generate a spurious link between the index and abundance. Such a link will cause the index to change when abundance changes even though the parameters of the underlying distribution of the individuals remain the same.

This study presents a statistical investigation of the properties of a number of different indices frequently used to study the relationship between distribution and abundance directly from samples consisting of counts of individuals. The indices are examined to determine whether they are statistically independent of the mean number of individuals per sample when the sample number shows either a Poisson or negative binomial distribution. When individuals are randomly distributed in space, the Poisson distribution is the appropriate distribution to describe the variation in the number of individuals per sample. However, when individuals are patchily distributed, the number of individuals per sample will follow a negative binomial distribution where the aggregation of individuals is reflected in the size parameter \( k \) (Lloyd 1967; Taylor et al. 1988). The indices and the aspect of distribution they reflect are listed in Table 1. Among the numerous published indices of spatial distribution available, we include only those that are frequently found in the literature and have been used to infer abundance-distribution relationships directly from samples (without prior smoothing or transformation). The properties of these indices are investigated both analytically and by simulation (Table 1). Further, the ability of the indices to detect changes in distribution is investigated in situations where the requirements of the indices with respect to statistical properties and spatial independence of samples are violated.

To illustrate the practical implications of using biased indices, we estimate the relationship between abundance and distribution of juvenile North Sea cod (Gadus morhua).

### Distribution indices investigated

#### Area occupied

The proportion of an area that is occupied by a given species is often determined from samples that each contain a number of individuals. The occupied proportion is assumed to be equal to the proportion of samples that contain counts of at least one individual (occurrence) or alternatively as the proportion of samples containing more than a fixed number of individuals. However, both of these indices are intrinsically linked to the mean number of individuals per sample when individuals are randomly distributed in space (Murawska and Finn 1988; Hartley 1998; Swain and Sinclair 1994) (Table 2) or in some proportion of space (Sileshi et al. 2009). Comparing occurrence with mean number of individuals in non-empty samples only (Macpherson 1989; Eycott et al. 2006; Webb et al. 2007) does not eliminate the link (Wright 1991). Occasionally, empty samples occur through a combination of random effects and sampling in truly empty areas (structurally empty samples). In these cases, the relative importance of the two types of empty samples can be estimated using zero inflated Poisson and negative binomial distributions (Silesi et al. 2009).

Assume that the data consist of individual counts in a number of samples, \( N \), of which a proportion \( p_c \) are taken in areas uninhabited because of, for example, adverse habitat conditions or contraction of the population. If the individuals within the inhabited area are distributed according to a negative binomial distribution with a parameter \( m \) reflecting the mean of individuals in a sample, the probability density function describing the probability of obtaining \( n \) individuals in a sample is

\[
P_{\text{Negbin}}(n) = \begin{cases} 
 p_c + (1 - p_c) \left( \frac{k}{k + m} \right)^k & \text{if } n = 0 \\
 (1 - p_c) \frac{\Gamma(n + k)}{\Gamma(n) \Gamma(k)} \left( \frac{k}{k + m} \right)^k \left( \frac{m}{k + m} \right)^n & \text{if } n > 0
\end{cases}
\]

where \( k \) is the size parameter of the negative binomial distribution. The parameters \( p_c, k, \) and \( m \) can be estimated using the R package zeroifnl (R Development Core Team 2010) or by maximizing the likelihood function.

#### Aggregation

#### Lloyds index of patchiness

Lloyd's index of patchiness is a function of the empirical mean number of individuals per sample \( \bar{n} \) and the sampling variance \( s^2 \) (Lloyd 1967):

\[
I_{LP} = 1 + \frac{s^2}{\bar{n}^2} - \frac{1}{\bar{n}}
\]

Unlike the proportion of non-empty samples, Lloyd's index is independent of the mean number of individuals per sample when applied to Poisson or negative binomially distributed data (Lloyd 1967). Lloyds index of patchiness can be regarded as an empirical estimate of the true index \( 1 + \frac{\sigma^2}{\bar{n}} - \frac{1}{\bar{n}} \) where \( m \) and \( \sigma^2 \) are the true mean and variance, respectively, of the distribution. The index estimated from the empirical mean and variance converges to the true index when the number of samples increases towards infinity. The true index takes the value 1 when the number of individuals in a sample is Poisson distributed and takes the value \( 1 + 1/k \) when data are negative binomially distributed, where \( k \) is the size parameter. In these cases, Lloyds index is approximately independent of the mean (Lloyd 1967) and can hence be used to compare patchiness of distributions with different means. However, this property is not retained when the distribution...
Table 1. Overview of the measures investigated.

<table>
<thead>
<tr>
<th>Type</th>
<th>Measure</th>
<th>Method used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupied area</td>
<td>Proportion of empty samples</td>
<td>Analytical derivation</td>
</tr>
<tr>
<td></td>
<td>Proportion of samples containing more than a fixed number of individuals</td>
<td>Analytical derivation</td>
</tr>
<tr>
<td></td>
<td>Proportion of structurally empty samples</td>
<td>Simulation(^a) and subsequent estimation</td>
</tr>
<tr>
<td></td>
<td>Lloyds index of patchiness(^a)</td>
<td>Simulation(^a) and subsequent estimation</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Measures based on the area below the Lorenz curve (Gini and A)</td>
<td>Simulation(^b) and subsequent estimation</td>
</tr>
<tr>
<td></td>
<td>Measures based on points on the Lorenz curve ((D_y))</td>
<td>Simulation(^b) and subsequent estimation</td>
</tr>
<tr>
<td>Geographical spread</td>
<td>Average distance to centre of gravity Area of the contour ellipse</td>
<td>Analytical supplemented by simulation(^c) to determine accuracy</td>
</tr>
</tbody>
</table>

- \(^a\)Lloyds index of patchiness converges to 1 + 1/k when data are negative binomially distributed, where k is the size parameter in the distribution.
- \(^b\)Simulations varied in number of samples simulated (20 to 500), the mean number of individuals per sample (0.1–10), and the distribution across samples to ensure that the greatest possible resolution was obtained in the region with significant bias. Simulations did not cover parameter combinations with constant bias (high densities) or for which bias could not be estimated (very low densities).

Table 2. The relationship between mean \(m\) and two indicators of aggregation (proportion of empty samples and Lloyds index of patchiness) for a range of statistical distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Variance</th>
<th>Proportion of zeros</th>
<th>Lloyds index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>(m(1 - \frac{m}{n}))</td>
<td>((1 - \frac{m}{n}))</td>
<td>1 - (\frac{m}{n})</td>
</tr>
<tr>
<td>Poisson</td>
<td>(m)</td>
<td>(e^{-m})</td>
<td>1</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>(m + \frac{e^m}{n})</td>
<td>(\frac{1}{1 + \frac{m}{n}})</td>
<td>1 + (\frac{m}{n})</td>
</tr>
<tr>
<td>Geometric</td>
<td>(m(m-1))</td>
<td>(\frac{1}{m})</td>
<td>(2(1 - \frac{1}{m}))</td>
</tr>
<tr>
<td>Normal(^b)</td>
<td>(\sigma^2)</td>
<td>0</td>
<td>(1 + \frac{\sigma^2}{m^2} - \frac{1}{m})</td>
</tr>
<tr>
<td>Log-normal(^b)</td>
<td>(m^2(e^{\sigma^2} - 1))</td>
<td>0</td>
<td>(e^{\sigma^2} - \frac{1}{m})</td>
</tr>
<tr>
<td>Gamma(^b)</td>
<td>(ms)</td>
<td>0</td>
<td>(1 + \frac{s}{m}(s - 1))</td>
</tr>
<tr>
<td>Exponential</td>
<td>(m^2)</td>
<td>0</td>
<td>2 - (\frac{m}{s})</td>
</tr>
</tbody>
</table>

\(^a\)\(\sigma^2\) denotes the variance.
\(^b\)\(e^{\sigma^2}\) denotes the log variance.
\(^c\)s denotes the scale parameter.

differs from that of a binomial, Poisson, or negative binomial distribution (Table 2). For individuals distributed according to the continuous lognormal and gamma distributions, Lloyds index depends on the mean, except in the special case where the gamma distribution has the scale parameter 1 (Table 2).

Lorenz curves

A Lorenz curve is the cumulative distribution of the samples ordered by ascending size. Indices based on Lorenz curves include the minimum area containing a fixed percentage of the population, \(D_x\), and the Gini index (Poulin 1993; Swain and Sinclair 1994; Woillez et al. 2007, among others). The indices describe aggregation as the difference between the observed distribution and a distribution where every sample contains the same number of individuals.

The measures \(D_{50}, D_{75}, D_{90},\) and \(D_{95}\), together denoted \(D_y\), are defined as the minimum area containing \(y\)% of the individuals and is determined by a point on the Lorenz curve (Fig. 1). \(D_y\) can be estimated by

\[
D_y = \frac{\sum_{i=1}^{N} n_i (z(l_i) - z(N))}{N}
\]

where \(z(l_i) = \sum_{j=1}^{i} n_{ij}/\sum_{j=1}^{n} n_{ij}\), \(l_i\) fulfill that \(z(l_i) \leq 1 - y/100 \leq z(l_i + 1)\), \(N\) is the number of observations, and \(n_{ij}\), \(n_{i2}, \ldots, n_{IN}\) are the observations in descending order. Hence, \(z(l_i)\) is the proportion of all individuals recorded in the smallest \(l_i\) samples. It is assumed here that the observations represent equally sized areas. \(D_y\) takes values between zero and \(y/100\). If \(D_y\) is zero or close to zero, the individuals are concentrated in a few samples. Conversely, if \(D_y = y/100\), all samples take equal values. Hence, increasing aggregation results in a decrease in \(D_y\). The indices \(D_{50}, D_{75}, D_{90},\) and \(D_{95}\) were introduced by Swain and Sinclair (1994) and have subsequently been reintroduced by Atkinson et al. (1997) and Anderson and Gregory (2000).

Another commonly used measure based on Lorenz curves is the Gini index, \(G\), defined as twice the area between the Lorenz curve and the diagonal (Fig. 1). The index is derived from socioeconomics and ranges from 0 to 1. It takes the value 0 when all samples are equal and 1 when all individuals are recorded in a single sample. It is estimated as

\[
G = \frac{1}{N - 1} \sum_{i=1}^{N} (N - i) [n_{i+1} - n_{i}]
\]

where \(N\) is the number of observations, and \(n_{1}, n_{2}, \ldots, n_{N}\) are the observations in ascending order. As above, the
area of the contour ellipse are heavily dependent on the distance to centre. The population would lead to a decrease in the average distance to centre and the area of the contour ellipse can be used directly as an estimate of the area containing a certain percentage of the population. The area can be calculated from the average squared distance to the centre and the covariance between the distance in latitudinal and longitudinal directions (as shown in Appendix A). Clearly, $d$, $s^2$, and the area of the contour ellipse are heavily dependent on the distributions of sampling sites. If sampling at a given time is only conducted in, say, the southern half of the usual sampling area, the average distance to centre and the area of the contour ellipse will both decrease even when no change in distribution occurs.

**Materials and methods**

**Area occupied**

The proportion of empty samples as a function of mean sample sizes when no areas were structurally empty was estimated directly from the relationships in Table 2. To investigate the case where a significant proportion of the samples are taken in structurally empty areas, accuracy and bias of the estimate of $p_e$ was investigated by simulating binomial distributed samples with a probability of 0.25 of being structurally empty ($p_e$). For the nonstructurally empty samples, number of individuals in each sample was simulated as negatively binomial distributed observations with a given size and mean. For each number of samples (250 and 500), mean (1, 2, ..., 10), and $k$ (0.5, 1 and 2), 1000 data sets were simulated, and the mean and the 2.5% and 97.5% quartiles of $\hat{p}_e$ were estimated using maximum likelihood estimation in R (R Development Core Team 2010). The number of samples was initially set at lower values to allow direct comparison with the other measures, but this resulted in estimates of 2.5% and 97.5% quartiles $\hat{p}_e$, which included the full range (0–1), and hence the number of samples had to be increased.
Aggregation

Lloyds index of patchiness

Lloyds index can either be estimated from moment estimates (eq. 2) or from maximum likelihood estimates of \( k \) (Lloyds index \( = 1 + 1/k \)). We used simulations to estimate the bias of the two different estimates at different number of samples (20, 40, …, 100), mean (0.1, 0.3, …, 0.9, 1.0, 1.5, …, 3.0), and \( k \) (0.1, 0.2, 0.5, and 1). First, we simulated negative binomial distributed samples at all combinations of number of samples, mean, and \( k \). We then used these data to estimate Lloyds index from moments and maximum likelihood estimates of \( k \). 1000 data sets were generated for each combination and the median, 2.5%, and 97.5% quantiles of the estimates estimated for each method. A number of the simulations resulted in a total number of individuals < 2 in the data set, and the parameters were therefore not estimable. These data sets were discarded. This occurred in a maximum of 30% of all cases.

Lorenz curves

We used simulations to estimate the indices for data from different distributions and parameter combinations. We restricted the analyses to statistical distributions of counts only and used a constant density in all areas and then sampled data assuming a Poisson or negative binomial distribution. A total of 10 000 samples were simulated at each average density. Average density varied between 0.1 and 10, and \( k \) varied between 0.1 and 1.

Geographical spread

To investigate the bias and precision of the indices of geographic range, we considered the estimate in one dimension with no loss of generality. The centre of gravity, \( C \), the average distance, \( d \), and the area of the contour ellipse (equivalent in the one-dimensional case to the square root of the average squared distance to centre) estimated from samples can be shown analytically to be unbiased or approximately unbiased, and the geographical spread of the underlying distribution does not affect their coefficient of variation, CV (Appendix A). The effect of varying mean density, standard deviation of the distribution of the underlying spatial distribution, and number of sampling locations on the uncertainty of \( d \) and \( s^2 \) estimated from samples were investigated using simulations. In these analyses, density was assumed to follow a normal distribution in the interval \( I = (-10; 10) \) with a centre of gravity equal to zero. Average density was changed by multiplying densities throughout the interval by a constant. Density \( \lambda(x) \) at each location \( x \) in the interval was calculated at \( n \) equidistant locations, and observations \( Y(x_1) \ldots Y(x_n) \) were generated assuming \( Y(x) \) to be Poisson distributed with mean \( \lambda(x) \); \( d \) and \( s^2 \) were then calculated using the generated observations. This was repeated 1000 times and the standard deviation of the indices estimated. Three values of geographical spread (standard deviation in the underlying normal distribution = 1, 2, and 3) and sampling intensity (number of samples = 20, 50, and 100) were investigated at mean densities ranging from 1 to 10.

Failing to meet assumptions

The distribution of individuals in space is rarely random with a common mean in all areas. To investigate the effect of a more ecologically likely spatial distribution, we investigated the ability of Lloyds index, the area of the contour ellipse, \( D_{95} \), and the Gini index to detect changes in aggregation of individuals with a bimodal spatial distribution (Fig. 2). We estimated the indices based on 500 simulated surveys sampling at 300 equidistant points along the distribution. Samples were drawn from a negative binomial distribution with a spatially varying mean described by the bimodal distribution and a common size parameter \( k = 0.4 \). Two different effects of a decrease in abundance were investigated: proportional decrease in all areas (proportional density model; Hilborn and Walters 1992; Shepherd and Litvak 2004) and disproportional decrease in low abundance areas (constant density and basin model; Hilborn and Walters 1992; Shepherd and Litvak 2004). Furthermore, the effect on the indicators of increasing the level of aggregation of the underlying spatial distribution was investigated. The level of aggregation of the underlying spatial distribution (Fig. 2) was quantified by estimating \( D_{95} \) of the underlying (true) distribution. \( D_{95} \) is an unbiased index of aggregation in this case, as the distributions are continuous normal distributions unaffected by sampling error (Swain and Sinclair 1994) rather than the number of individuals observed in samples. To be a reliable index, there must be a tight relationship between the index and \( D_{95} \), and index values obtained at other average densities should fall inside the quantiles (corresponding to no falsely detected changes in distribution).

Data example: North Sea cod

The data used to illustrate the problems resulting from using biased indices consisted of survey catch rates of 1-year-old cod derived from catches in the International Council for the Exploration of the Sea (ICES) International Bottom Trawl Survey (ICES International Bottom Trawl Survey Database, February 1983 to 2010, http://www.ices.dk). Haul duration generally varies between 0.5 and 1 h, but to assure that catch rates were comparable, only hauls
lasting between 25 and 35 min were used in this study. The survey follows a stratified-random design with two hauls taken within each ICES statistical rectangle (0.5° latitude × 1° longitude) on average. Catch of 1-year-old cod was estimated by applying an age–length key to the observed catch at length. If the total number of fish caught of all ages is negatively binomial distributed, multiplying this number by a fraction results in a number that is not strictly discretely distributed. In our study, we focused on 1-year-olds, as their length distribution is generally well separated from older fish, and hence the assumption of observing an integer is reasonable. There were, however, cases where a caught fish was of a length with a significant probability of being age 2. In these cases, we rounded the number caught to the nearest integer. We believe that the error introduced by this is small compared with the variation we would introduce by examining the total (integer) number of cod caught of all ages.

From catches of 1-year-old cod, we estimated mean catch rate, Lloyds index, the proportion of empty samples, $D_95$, the Gini index, average distance to centre, and the area of the contour ellipse ($P = 0.40$) for every year. The proportion of structurally empty samples was not estimated because of the low accuracy at low values of $k$. For the two geographic indices, catch rates were averaged within statistical rectangles prior to estimation to avoid bias due to changes in sampling intensity. Finally, a model of the form $\ln(\text{measure}) = a + b \times \ln(\text{mean catch})$ was fitted, and the statistical test of $b = 0$ was performed.

Results

Area occupied

The intrinsic link between mean sample size and the proportion of nonzero samples diminishes rapidly with increasing mean if individuals are Poisson distributed. At a mean sample size of 5, the proportion of nonzero samples is greater than 0.99 (Table 2, Fig. 3). However, if data are negatively binomial distributed, the problem caused by "false" empty samples alone is aggravated (Fig. 3). Hence, conclusions regarding distribution can only be based directly on the occurrence of nonzero samples when the mean number of individuals per sample is large (>4 for Poisson distributed data, >10 for negative binomial distributed data), and the probability of obtaining zero samples by random effects alone is small.

The maximum likelihood method provides unbiased estimates of $p_e$ (bias on average 2% with no trend), which are reasonably precise for mean sample sizes above 5 (Fig. 4). The precision of the method relies on both mean and size parameter of the negative binomial distribution. Hence, distributions with low mean or a size parameter below 1 are unlikely to provide precise estimates of $p_e$ unless the number of samples is very large.

Aggregation

Lloyds index of patchiness

The moment estimate of Lloyds index of patchiness was consistently biased at high index values (>2), even in the 200 samples case (Fig. 5). In contrast, the bias in the maximum likelihood estimate was small when more than 40 samples were taken and the mean exceeded 0.5 (Fig. 5). When the number of samples exceeded 100, the maximum likelihood estimate was approximately unbiased for all means. The variation in both estimates was large when Lloyds index was large or mean was small, but consistently larger for moment estimates than for maximum likelihood estimates (Appendix A). When setting the acceptable upper limit for the CV of the estimate at 20%, none of the two methods are sufficiently accurate to allow the estimation of Lloyds index from less than 60 samples at any value of $k$ (Appendix A). Hence, to estimate Lloyds index, it is a requirement that the number of samples taken is large, and this is increasingly important as either the mean or $k$ decreases. Further, it is preferable to use maximum likelihood estimates of $k$ rather than moments to estimate the index.

Lorenz curves

$D_95$ and the Gini index are both highly dependent on the mean at low mean numbers of individuals in a sample (Fig. 6). If the distribution of individuals is known without error, $D_95$ or the Gini index can obviously be calculated without bias using eqs. 3 and 4, respectively. However, this is rarely the case, and when observations are subject to sampling variation, indices based on Lorenz curves are systematically biased when the mean number of individuals per sample is below 4 (Fig. 6).

Geographical spread

Though both indices are unbiased, the CV of both $d$ and $s^2$ increases as average density is decreased (Fig. 7). The effect is greatest at average densities below 5 but levels off at average densities between 5 and 10. As expected, increasing the number of samples taken decreases the CV of the indices, and with a sampling intensity of 100 samples, the CV of $d$ and $s^2$ remains below 0.08 and 0.15, respectively, even at an average density of 1. The spread of the underlying distribution does not affect the CV of $d$ and $s^2$ (Fig. 7). However, their precision depends heavily on sampling intensity and
mean density, and the CV increases as average density or sampling intensity is decreased (Fig. 7).

**Failing to meet assumptions**

In spite of the fact that the data were not derived from a negative binomial distribution, the maximum likelihood estimate of Lloyds index reflected both the direction and level of change in distribution accurately (Fig. 8). There was a tight relationship between Lloyds index and $D_{95}$ of the underlying distribution, indicating that these indices reflect the same aspect of the distribution (Fig. 8). Though both $D_{95}$ and the Gini index produced values with a tight relationship with the underlying distribution, their actual values were far from that of the underlying distribution (Figs. 8b and 8c). Furthermore, varying only density produced indices outside the quantiles though no change in distribution occurred (Fig. 8). The proportional density model ($D_{95} = 0.74$ at different means) produced values of Lloyds index that did not differ significantly, confirming the absence of bias in this measure. The difference between the distributions caused no change in ellipse area, and this indicator did not reflect any of the changes in the bimodal distribution. Hence, only Lloyds index fulfilled both of the requirements to the index, as it reflected distributional change and was robust to changes in abundance in the absence of changes in distribution.

An overview of the indices investigated and their ability to provide unbiased estimates of the relationship between abundance and distribution and their sensitivity to violations of the assumptions is provided (Table 3).

**Example: North Sea cod**

The conclusions reached differed widely depending on whether biased or unbiased indices were used (Fig. 9). If the proportion of empty samples or $D_{95}$ was used as the indicator of the distribution, one would draw the (erroneous) conclusion that the area occupied is decreasing as abundance decreases (Figs. 9d and 9f; $P < 0.01$ in both cases). In contrast, there is no significant relationship between abundance and aggregation or between abundance and geographic extent when the unbiased indices are used (Figs. 9a, 9b, and 9c; $P > 0.05$ in all cases). The Gini index showed no significant relationship with the mean (Fig. 9e; $P = 0.3238$) and would thus not have produced a false positive in this case. Though the relationship between the mean of the two was not significant, variation in both Lloyds index and average distance to centre appeared to increase as mean catch decreased.
**Discussion**

We have shown here that a number of frequently used indices of distribution are intrinsically linked to the mean sample size when the mean is small. Using such indices may lead to the conclusion that a change in distribution has occurred when in fact there has only been a change in bias. Apart from the effect that this has on the understanding of the ecology of individual species, it may lead to inefficient conservation of natural populations if conservation management is erroneously based on assumptions of aggregation at low stock size. The problem can be avoided by using the proportion of structurally empty areas, Lloyds index, and the area of the contour ellipse as indices of area occupied, aggregation,
Fig. 7. Investigating the effect of sampling on the bias of the measured mean squared distance to centre \( s^2 \). Data include median bias ((estimated index – true index)/true index; middle line) and 2.5% quantile (bottom line) and 97.5% quantile (top line) of bias. (a) Effect of spread of the underlying distribution: standard deviation = 1 (solid line), 2 (dashed line), and 3 (dotted). Number of samples = 50 in all cases. (b) Effect of number of sampling sites along the transect: 20 (solid), 50 (hatched) and 100 (dotted). Standard deviation of the underlying distribution = 2 in all cases.

Fig. 8. Relationship between \( D_{95} \) of the underlying distribution and Lloyds index (a), \( D_{95} \) (b), Gini index (c), and the area of the contour ellipse (d) estimated from simulated samples of the population. Diamonds, estimated mean index; dashed line lines, 2.5% and 97.5% quantiles of the estimated index; solid line, power function describing the relationship between the two; dash-dotted line, values of the underlying (true) distribution (\( D_{95} \) and Gini index only). All solid symbols depict indices estimated at an average density of 4.9, while open symbols are derived from average densities varying between 1.2 and 9.9 at a constant \( D_{95} \) of 0.74.
and geographical spread, respectively. However, although these indices are unbiased, they all showed consistent increases in variance at low mean sample size. Furthermore, the proportion of structurally empty areas was effectively impossible to estimate at values of \( k \) generally observed in analyses of marine fish (Murawski and Finn 1988; this study) unless the number of samples was very large and the area of the ellipse only revealed changes in unimodal distributions. Lloyds index was remarkably robust to changes in the underlying distribution and would appear to be the best choice for low-mean data. It furthermore also reflected modest changes in occupied area.

Lloyds index is robust to deviations from a negative binomial distribution. However, two conditions must be fulfilled:

1. A large number of samples, mean > 3 and \( k > 1 \)
2. Poisson or negative binomial distributed data

### Table 3. Overview of the bias of all measures and requirements of unbiased measures.

<table>
<thead>
<tr>
<th>Type</th>
<th>Measure</th>
<th>Biased?</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupied area</td>
<td>Proportion of empty samples</td>
<td>Approximately unbiased when mean &gt; 5</td>
<td>Large number of samples, mean &gt; 3 and ( k &gt; 1 )</td>
</tr>
<tr>
<td></td>
<td>Proportion of structurally empty samples</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proportion of samples containing more than a fixed number of individuals</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Aggregation</td>
<td>Lloyds index of patchiness (moment estimate)</td>
<td>Yes</td>
<td>Poisson or negative binomial distributed data</td>
</tr>
<tr>
<td></td>
<td>Lloyds index of patchiness (maximum likelihood estimate)</td>
<td>Approximately unbiased when number of samples &gt; 100</td>
<td>Poisson or negative binomial distributed data</td>
</tr>
<tr>
<td></td>
<td>Measures based on points or area below the Lorenz curve (( D_0, Gini, and A ))</td>
<td>Approximately unbiased when mean &gt; 4</td>
<td></td>
</tr>
<tr>
<td>Geographical spread</td>
<td>Average distance to centre of gravity</td>
<td>No but CV large when number of samples &lt; 100</td>
<td>Constant area sampled and unimodal spatial distribution</td>
</tr>
<tr>
<td></td>
<td>Area of the contour ellipse</td>
<td>No but CV large when number of samples &lt; 100</td>
<td>Constant area sampled and unimodal spatial distribution</td>
</tr>
</tbody>
</table>

### Fig. 9. Relationship between abundance and distribution of juvenile Atlantic cod in the North Sea using unbiased (a, b, and c) and biased (d, e, and f) measures of distribution.
for the estimator to remain unbiased. Firstly, estimating the common form of Lloyds index on non-empty samples only (Hewitt 1981; Stabeno et al. 1996) provides biased results (Bradbury et al. 2003), and eq. 1 or Lloyds index for positive samples ($k_+$; Lloyd 1967) should be used instead if there are structurally empty samples. Secondly, the observations used must be counts rather than concentrations. Concentrations are often used, in particular when the sampled volume is unequal between samples (McGurk 1986; Stoffels et al. 2003; Bez 2000). However, the derived concentrations are not negative binomial distributed counts, and Lloyds index is therefore no longer independent of the mean (Appendix A). Furthermore, the problems with small counts are retained when numbers are divided by, for example, volume of water filtered. This means that the bias in $D_{95}$ and the Gini index is retained (Appendix A).

To justify the use of the proportion of non-empty samples as a measure of occupied area, two statements are frequently made: The first claim is that it is impossible to distinguish true zeros from false (Hewitt 1981; Hanski et al. 1993; Gaston et al. 1998). It is clear from this and other studies (Lloyd 1967; MacKenzie et al. 2003; Sileshi et al. 2009) that it is indeed possible to estimate the proportion of true zeros in many cases. The second is that for an area to be occupied, there must be at least one individual present in the samples. However, if a species is wide ranging, samples with fewer than one individual will occur by chance when one or more individuals spend only part of their time in the area or when the probability of detecting an individual is less than 1. It is clear that if there are only a few specimens left of the species, these few specimens are unlikely to inhabit the entire area. However, the question is whether the few remaining specimens occupy all areas but rather whether we would be able to detect a change in their distribution. In the case where we observe one or two specimens, their distribution will provide us with virtually no information of the distribution of the population. Hence, our conclusion in this case must be that the species is rare and that the current distribution cannot be evaluated because of the low number of observations. The result of the general acceptance of the two statements is that the intrinsic correlation between the abundance and proportion of non-empty samples is often ignored (Planque and Fromentin 1996; Blackburn et al. 1998; Freckleton et al. 2006, among others), though some authors have tried to avoid the problem by comparing occurrence with mean number of individuals in occupied samples only (Macpherson 1989; Eycott et al. 2006; Webb et al. 2007). The latter procedure does not eliminate the problem (Wright 1991) though it does decrease the correlation between abundance and occupancy considerably (Wilson 2008, 2011), in some cases leading to negative relationships (Blackburn et al. 2006; Wilson 2011).

Biased distribution indices have been used in studies in a wide number of ecological groups, including parasites (Poulin 1993), plants (Eycott et al. 2006), invertebrates (Foggo et al. 2003), fish (Blanchard et al. 2005), birds (Zuckerberg et al. 2009), and mammals (Holt and Gaston 2003), and it is likely that they have seriously affected the conclusions in a number of cases. For example, Winters and Wheeler (1985) and Saville and Bailey (1980) used the area fished by commercial fishermen as a measure of distribution, a measure that is likely to be equivalent to the area in which a catch rate higher than some minimum level can be obtained. Crecco and Overholtz (1990) used the proportion of the area where survey catch rate was above a fixed level. However, as the proportion of samples that contain more than a fixed catch is linked directly to mean density, the relationship between abundance and distribution cannot be derived from these studies. Turning to indices based on Lorenz curves, studies of species showing low mean catch rates or greatly skewed data have demonstrated significant relationships between abundance and both $D_{95}$ (Fisher and Frank 2004; Blanchard et al. 2005) and the Gini index (Laurel at al. 2004) of the type shown herein.

The example of cod in the North Sea illustrates the effects of using biased indices to estimate abundance–distribution relationships. Had we used only the proportion of empty samples and $D_{95}$, we would have concluded that cod concentrated in a few areas as density declined. However, this conclusion is caused by the use of biased indices rather than by changes in distribution as is evident when we use unbiased indices to investigate the abundance–distribution relationship. Even if we had used unbiased indices, we would have been tempted to conclude that though cod did not aggregate, they exhibited a greater variability in aggregation behaviour at low density. However, it is clear from the simulation studies that this type of increase in variance is to be expected from the decreased accuracy of the estimator even in the absence of a change in aggregation. The lack of change in aggregation of juvenile cod as density declines is in accordance with a study of this species in Placentia Bay, which found no significant difference in $D$ (Robichaud and Rose 2006). Similarly, Murawski and Finn (1988) found no significant correlation between average distance to centre and abundance of age 0 and 1+ on Georges Bank. Older cod off the east coast of Canada showed pronounced aggregation as density declined, and the proportion of high catch rates remained virtually constant as abundance declined (Hutchings 1996). Similar aggregative relationships have been found for southern Gulf of St. Lawrence cod age 3 and older (Swain and Sinclair 1994). Determining whether these differences in aggregative behaviour are caused by ontogenetic shifts in behaviour or by differences between stocks will require more detailed analyses of unbiased abundance–distribution relationships.

In conclusion, we recommend that the frequency distribution and mean of the samples is examined in detail before deciding on an unbiased measure of distribution. The proportion of empty samples should not be used to estimate abundance – occupied area relationships in any case. Instead, the proportion of structurally empty samples can be used to estimate empty areas except in cases where the data are highly skewed. In these cases, it is not possible to estimate the proportion of structurally empty samples unless the number of samples is very large (>500), and even in this case, mean sample sizes < 1 are unlikely to provide reliable results. If the mean number of individuals in the samples is >4, $D_{95}$ or the Gini index should provide approximately unbiased results of the relationship between abundance and aggregation. Both indices are furthermore unbiased if applied to smoothed data derived from a spatial model (Swain and Sinclair 1994; Lewy and Kristensen 2009). The average distance to centre, the average squared distance to centre, and the area...
of the ellipse should only be used as indices of geographical spread when the spatial distribution is strictly unimodal, as these indices are insensitive to changes in other distributions. For indices of aggregation based on count data, Lloyds index estimated using maximum likelihood methods is the preferable choice, possibly supplemented by the area of the ellipse to indicate geographical spread when the distribution is unimodal. Both indices are unbiased down to very low mean samples sizes and suffer only from increased variation as mean sample size is reduced.

Acknowledgements

The work of A. Rindorf was funded by the SUNFISH project granted by the Danish agency for Science, Technology and Innovation. The work of P. Lewy was funded by the FP7 project VECTORS (project No. 266 445).

References


Appendix A

Bias in Lloyds index estimated by moments and maximum likelihood

Figures A1 and A2 show bias and accuracy of estimates of Lloyds index of patchiness based on moments (eq. 2; Fig. A1) or maximum likelihood estimates of k (Fig. A2).

Estimating the area of the contour ellipse and bias of C and d due to sampling error

The area of the contour ellipse A<sub>P</sub> containing p percent of the observations can be estimated as

\[
A_p = c_p \sqrt{s_x^2 + s_y^2 (1 - \rho^2)}
\]

where \(c_p = \chi^2(2p, \pi)\), and \(s_x, s_y\) are the standard deviations in the x and y directions where \(\rho\) is the correlation coefficient.

To compare distribution with abundance, only an index of geographical distribution is needed, and it is of no consequence what probability level is used (the \(\chi^2\) value, since this is simply a scaling factor. In the example given below, we have used \(c_p = 1\) without lack of generality. The area of the contour ellipse is independent of the mean number of individuals in a sample and of mean in the x and y directions because this is the case for both standard deviations and correlation.

To investigate whether the measures C and d and the area of the contour ellipse are unbiased, we consider the one-dimensional case for which \(\lambda(x)\) denotes the intensity of a species distributed on the x axis in a given interval, I. In this case, an unbiased estimate of \(s\) is equivalent to an unbiased estimate of the contour ellipse. The true centre of gravity \((C_{true})\) average distance to centre \((d_{true})\), and variance \((s_{true}^2)\) are

\[
C_{true} = \frac{\int x \lambda(x) dx}{\int \lambda(x) dx}
\]

\[
d_{true} = \frac{\int |x - C_{true}| \lambda(x) dx}{\int \lambda(x) dx}
\]

Published by NRC Research Press
Now assume that we have observations at $N$ locations and denote the density at location $x_i$ as $\lambda_i = \lambda(x_i)$.

The centre $C$, average distance to centre $d$, and variance $s^2$ based on the locations $x_1$ to $x_N$ are correspondingly estimated by

$$C = \frac{\sum x_i \lambda_i}{\lambda_*}$$

$$d = \frac{\sum |x_i - C| \lambda_i}{\lambda_*}$$

$$s^2 = \frac{\sum (x_i - C)^2 \lambda_i}{\lambda_* - 1}$$

where $\lambda_* = \sum \lambda_i$. We now assume that the observed densities $\eta_i$ include sampling noise.

The estimated centre, average distance to centre, and average squared distance to centre based on the observations $\eta_1$ to $\eta_n$ are then
Fig. A2. Bias and accuracy of estimates of Lloyds index of patchiness based on maximum likelihood estimates of $k$. Data for median bias ((estimated index – true index)/true index; solid line), 2.5% quantile and 97.5% quantile of bias (dashed lines) of 1000 simulations for each combination of number of samples, mean, and $k$ are shown. Horizontal dashed lines indicate 0.4, corresponding roughly to a coefficient of variation (CV) of the index of 20%.

$$\hat{C} = \frac{\sum x_i \eta_i}{\eta_*}$$

$$\hat{d} = \frac{\sum |x_i - \hat{C}| \eta_i}{\eta_*}$$

$$\hat{s}^2 = \frac{\sum (x_i - \hat{C})^2 \eta_i}{\eta_* - 1}$$

where $\eta_* = \sum \eta_i$. We assume that the observations for given densities are independent Poisson distributed variables, i.e., that

$$\eta_i | \lambda_i \sim \text{Poisson}(\lambda_i)$$

and hence $E(\eta_i) = \lambda_i$ for $i = 1, \ldots, N$.

For the Poisson case, it can be shown that $\hat{C}$, $\hat{d}$, and $\hat{s}^2$ are unbiased estimates of $C$, $d$, and $s$, respectively, i.e., $E(\hat{C}) = C$, $E(\hat{d}) = d$, and $E(\hat{s}^2) = s^2$.

Using the fact that the Poisson distributed variables $\eta_1$ ... $\eta_n$ for given sum $\sum \eta_i = \eta_*$ follows the multinomial distribution with the number of trials equal to $\eta_*$, probabilities $\frac{\eta_i}{\eta_*}$, and mean $E(\eta_i | \eta_*) = \eta_* \frac{\lambda_i}{\eta_*}$, we find that
Lloyd's index, $D_{S5}$, and the Gini index from concentrations

To demonstrate that the bias in $D_{S5}$ and Gini is retained even when transforming count data to concentrations by dividing with an uneven volume and that this also biases Lloyd's index, a simulation study was performed to resemble an experiment with 25 samples of 1 L, 50 samples of 2 L, 50 samples of 3 L, and 25 samples of 4 L. The number of individuals in 1 L follows the negative binomial distribution with a fixed mean and size = 0.4. Hence, the number of copepods in 2 L was simulated as the sum of two simulations from a negative binomial distribution with mean $n$ and size 0.4 and so forth. Concentrations were then estimated by dividing the simulated number of individuals by the volume of the sample (2 L in this case). After simulating concentration in all 150 samples, Lloyd's index (moment estimate), $D_{S5}$, and the Gini index were estimated. The simulation and estimation was performed 500 times, and the mean, 2.5, and 97.5 quantiles of each indicator were estimated (Fig. A3). It is clear that all three indicators show a change in aggregation, though the patchiness of the underlying distribution remains unchanged.

Bias when estimating Lloyd's index, $D_{S5}$, and the Gini index from concentrations

To demonstrate that the bias in $D_{S5}$ and Gini is retained even when transforming count data to concentrations by dividing with an uneven volume and that this also biases

\[
E(\hat{C}) = E_{\eta_*}[E(\hat{C} | \eta_*)] = E_{\eta_*}[\frac{1}{\eta_*} \left( \sum x_i \eta_i | \eta_* \right)] = E_{\eta_*} \left( \frac{1}{\eta_*} \sum x_i \eta_i \frac{\lambda_i}{\lambda_*} \right) = \sum \frac{x_i \lambda_i}{\lambda_*} = C
\]

\[
E(\hat{d}) = E_{\eta_*}[E(\hat{d} | \eta_*)] = E_{\eta_*} \left[ \frac{1}{\eta_*} \left( \sum |x_i - \hat{C}| \eta_i | \eta_* \right) \right] = E_{\eta_*} \left( \frac{1}{\eta_*} \sum |x_i - \hat{C}| \eta_i \frac{\lambda_i}{\lambda_*} \right) = \frac{\sum |x_i - \hat{C}| \lambda_i}{\lambda_*} = d
\]

and

\[
E(\hat{s}^2) = E_{\eta_*}[E(\hat{s}^2 | \eta_*)] = E_{\eta_*} \left( \frac{1}{\eta_* - 1} \left[ \sum (x_i - \hat{C})^2 \eta_i | \eta_* \right] \right) = E_{\eta_*} \left( \frac{1}{\eta_* - 1} \sum (x_i - \hat{C})^2 \eta_i \frac{\lambda_i}{\lambda_*} \right) = \sum \frac{(x_i - \hat{C})^2 \lambda_i}{\lambda_* - 1} = s^2
\]

where $E_{\eta}$ denotes expectation with respect to the distribution function of $\eta_*$. This proves the statement for the Poisson case. For any other distribution for which $E(\eta_i) = \lambda_i$, the three quantities are at least approximately unbiased, as both the numerators and the denominators of the quantities are unbiased.

As the number of observations tends towards infinity, $E(\hat{C})$, $E(\hat{d})$, and $E(\hat{s}^2)$ tend towards the true values $C_{true}$, $d_{true}$, and $s_{true}^2$, respectively.