



## Assessment of species diversity from species abundance distributions at different localities

Steinar Engen, Bernt-Erik Sæther, Anne Sverdrup-Thygeson, Vidar Grøtan and Frode Ødegaard

*S. Engen, Centre for Conservation Biology, Dept of Mathematical Sciences, Norwegian Univ. of Science and Technology, NO-7491 Trondheim, Norway. – B.-E. Sæther (bernt-erik.sather@bio.ntnu.no) and V. Grøtan, Centre for Conservation Biology, Dept of Biology, Norwegian Univ. of Science and Technology, NO-7491 Trondheim, Norway. – A. Sverdrup-Thygeson and Frode Ødegaard, Norwegian Inst. for Nature Research, Tungasletta 2, NO-7047 Trondheim, Norway.*

We show how the spatial structure of species diversity can be analyzed using the correlation between the log abundances of the species in the communities, assuming that two communities at different localities can be described by a bivariate lognormal species abundance distribution. A useful property of this approach is that the log abundances of the species at two localities can be considered as samples from a bivariate normal distribution defined by only five parameters. The variances and the correlation can be estimated by maximum likelihood methods even if there is no information about the sampling intensity and the number of unobserved species. This method also enables estimation of over-dispersion in the sampling relative to a Poisson distribution that allows sampling adjustment of the estimate of  $\beta$ -diversity. Furthermore, we also obtain a partitioning of species diversity into additive components of  $\alpha$ -,  $\beta$ - and  $\gamma$ -diversity. For instance, if the correlation between the log abundances of the species is close to one, the same species will be common and rare in the two communities and the  $\beta$ -diversity will be low. We illustrate this approach by analysing similarities of communities of rare and endangered species of oak-living beetles in south-eastern Norway. The number of recorded species was estimated to be only 48.1% of the total number of species actually present in these communities. The correlations among communities dropped rather quickly with distance with a scaling of order 200 km. This illustrates large spatial heterogeneity in species composition, which should be accounted for in the design of schemes of such series for assessing species diversity in these habitat-types.

A proper assessment of species diversity in an area is important for examining many relevant questions in ecology as well as for development of management actions for conserving biodiversity. In many cases the data are based on single samples of species abundances from different sampling sites. The species diversity is then estimated by a comparison of the similarity in species composition, e.g. by the use of the some indices of community similarity (Magurran 2004, Legendre et al. 2005, Anderson et al. 2006) or by some information theoretic measures (Levins 1968, Ludovisi and Taticchi 2006). A central problem in such assessments of species diversity is to account for sampling. In general, the number of species recorded will be closely related to the sampling effort (Pielou 1975, Lande 1996), i.e. the number of species in a sample will increase with the number of individuals sampled. As a consequence, the number of species recorded and hence the similarity of the communities will be strongly influenced by variation in sampling intensity (Chao et al. 2005). Basically, two different approaches have been suggested to account for the effects of sampling on estimates of species diversity. One approach is to assess sample size-effect by using non-parametric techniques such as rarefaction (Mao and Colwell

2005, Crist and Veech 2006). A problem with this approach is that the procedures chosen for standardization of the data sets can give very different results, and it is not always clear which measure of the species diversity that is more appropriate (Gotelli and Colwell 2001). The other set of approaches for assessing species diversity is to use a parametric species abundance model (Golicher et al. 2006). A disadvantage of this approach is that the estimates of many parameters can be sensitive to the choice of community model (Palmer 1990, Baltanás 1992, Magurran 2004, Williamson and Gaston 2005).

An important contribution to the study of species diversity was Whittaker's (1970, 1972) partitioning of species diversity into components due to  $\alpha$ -diversity within localities,  $\gamma$ -diversity in the whole region and  $\beta$ -diversity which he defined as turnover of species among samples at different localities. This decomposition resulted in a large number of studies to estimate the relative contribution of these components to species diversity (McGill et al. 2007). In particular,  $\beta$ -diversity received huge attention because this component embeds many of the fundamental processes such as depletion of species numbers or homogenization of species composition across localities that affects species

diversity locally as well as regionally. For instance, many studies covering a large number of taxa have studied how  $\beta$ -diversity varies along environmental gradients (Condit et al. 2002, Davidar et al. 2007, Novotny et al. 2007, Qian and Ricklefs 2007) or in relation to different environmental modifications (Melis et al. 2007), which are often induced by various forms of human activities. Furthermore,  $\beta$ -diversity is also crucial in conservation planning or for management actions (Ferrier 2002, Rooney et al. 2007) because it strongly affects the necessary areas or the number of critical habitat types to protect for conserving species diversity.

Whittaker's (1970, 1972) original definition of  $\beta$ -diversity entailed a multiplicative partitioning, so that  $\beta = \gamma/\alpha$  is dimensionless. This precludes any direct comparisons of  $\beta$  for different values of  $\gamma$  and  $\alpha$ . In an important contribution Levins (1968) showed that species diversity also could be partitioned into additive components within and between communities. This approach was later extended by Lande (1996) and Crist and Veech (2006). In this paper we present a method for calculating the similarity of two communities at different sites that maintain the additive properties among the diversity components as well as taking into account sampling effects with possible over-dispersion relative to random Poisson sampling. We use a parametric approach, utilizing the results of Preston's (1948, 1962, 1980) surveys of a large number of communities showing that species abundance distributions often are approximately lognormally distributed (Tokeshi 1993). Adopting the dynamical model for communities of Engen and Lande (1996), Engen et al. (2002a), and Lande et al. (2003), we get the conceptually simple result that two communities at different locations can be described by a bivariate lognormal species abundance distribution. More precisely, if there are  $s$  species at these locations their log abundances  $(x_i, y_i)$ ,  $i = 1, 2, \dots, s$  can be considered as a sample of  $s$  independent observations from a bivariate normal distribution. This model is based on the assumption that each species has the Gompertz type of density-regulation and include possible heterogeneity defined by assuming that the growth rates are normally distributed among species. Further details on species area curves derived from this model and a more general discussion of heterogeneity is given by Engen (2007a, 2007b). Notice that, although all species by definition are present in both communities, the log abundance has no lower limit. Therefore, a species that is observable in one community may still have so small abundance in the other one that it can never be observed and is therefore in practice absent. The correlation  $\rho_{xy}$  between the communities can then be viewed as an index of similarity. If  $\rho_{xy}$  is close to one, the same species will have large and small abundances in the two communities, if  $\rho_{xy}$  is zero there is no relation between the ordering of species abundances, and if  $\rho_{xy}$  is negative, species that are abundant in one community are likely to be rare in the other. In this way, we can identify the similarity between communities in a region and use this to estimate Whittaker's (1972) components of diversity.

The bivariate normal distribution has five parameters, the means and variances of the two marginal normal distributions and the correlation. If a species with log abundance  $x$  and abundance  $e^x$  in one community has  $N_x$

representatives in the sample, one commonly assumes that the sampling can be described by the Poisson distribution as first proposed by Fisher et al. (1943). Formally, this means that  $N_x$  is Poisson distributed with mean  $v_x e^x$  when conditioned on  $x$ , where  $v_x$  is a measurement of the sampling intensity. When  $x$  is normally distributed among species with mean  $\mu_x$  and variance  $\sigma_x^2$ , the unconditional distribution of  $N_x$  (not conditioning on the abundance) is the Poisson lognormal distribution (Grundy 1951, Bulmer 1974, O'Hara 2005) with parameters  $(\ln v_x + \mu_x, \sigma_x^2)$ . Joint samples from two communities,  $(N_x, N_y)$  then follows a bivariate Poisson lognormal distribution (Engen et al. 2002a, Lande et al. 2003) with parameters  $(\ln v_x + \mu_x, \sigma_x^2, \ln v_y + \mu_y, \sigma_y^2, \rho_{xy})$ . These five parameters can be estimated by maximum likelihood from samples from the two communities. If the sampling intensities are unknown, the mean values  $\mu_x$  and  $\mu_y$  can not be estimated because they are confounded with  $\ln v_x$  and  $\ln v_y$ . However, for most practical purposes, one will usually be interested in the variance of the lognormal species abundance distribution and the correlation between two communities. Fortunately, these parameters can be estimated even if there are no available information about sampling intensities available.

In most cases the Poisson distribution is not a correct sampling model because, for a given abundance, one will usually find over-dispersion relative to the Poisson distribution. Thus, the variance of  $N_x$  will usually be much larger than the mean. Therefore, Engen et al. (2002a) proposed using the Poisson-lognormal as sampling distribution. For a given abundance  $x$ , the number of individuals observed is then Poisson with mean  $v_x Z e^x$ , where  $\ln Z$  is normally distributed with mean  $-\theta^2/2$  and variance  $\theta^2$  and takes independent values for each species sampled. Then  $Z$  is lognormally distributed with unit mean and variance increasing with  $\theta$ . The unconditional distribution of  $N_x$  will now be the Poisson lognormal with parameters  $(\ln v_x - \theta^2/2 + \mu_x, \sigma_x^2 + \theta^2)$  and the bivariate distribution of  $(N_x, N_y)$  is the bivariate Poisson lognormal with parameters  $(\ln v_x - \theta^2/2 + \mu_x, \sigma_x^2 + \theta^2, \ln v_y - \theta^2/2 + \mu_y, \sigma_y^2 + \theta^2, \rho)$ , where

$$\rho = \frac{\rho_{xy} \sigma_x \sigma_y}{\sqrt{(\sigma_x^2 + \theta^2)(\sigma_y^2 + \theta^2)}} \quad (1)$$

Hence, the correlation  $\rho$  estimated from parallel samples from the same community, having by definition  $\rho_{xy} = 1$ , will be smaller than one if there is over-dispersion in the sampling relative to the Poisson distribution. Accordingly, an estimated correlation different from one gives information about the magnitude of over-dispersion in the sampling defined by  $\theta$ .

In this paper we will demonstrate this bivariate lognormal approach for assessing species diversity in an area in which species-abundance distributions are available from several localities, using a sample of communities of oak-living beetles in southern Norway. We show how to calculate  $\alpha$ -,  $\beta$ -, and  $\gamma$ -diversity as well as important parameters such as the variance of the lognormal distribution, similarities among communities and estimates of the total number of species present. An important advantage of this approach is that it is based on specific assumptions that can be evaluated.

## The Poisson lognormal distribution

### Univariate Poisson lognormal model

A Poisson mixture is obtained by assuming that the mean  $\lambda$  of the Poisson distribution for each recording of the discrete variable  $N$  is generated from some distribution  $f(\lambda)$ , giving

$$p_n = P(N = n) = \int P(N = n|\lambda)f(\lambda)d\lambda$$

Writing  $\ln \lambda = \mu + \sigma u$  where  $u$  is a standard normal variate,  $\ln \lambda$  is normal with mean  $\mu$  and variance  $\sigma^2$  and  $\lambda$  is lognormally distributed with parameters  $\mu$  and  $\sigma^2$ . The mean value of  $\lambda$  is then  $e^{\mu + \sigma^2/2}$  while the squared coefficient of variation is  $e^{\sigma^2} - 1$ . The above unconditional distribution of  $N$  can now be written on the form

$$p_n = \frac{e^{n\mu - e^{\mu}}}{n!} \int_{-\infty}^{\infty} e^{n\sigma u - e^{-\sigma u}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

This is the Poisson lognormal distribution with parameters  $\mu$  and  $\sigma^2$  first introduced in biology by Grundy (1951) and as a species abundance distribution by Bulmer (1974). When applied to describe the distribution of observed abundance in a community, we assume that the log abundance  $x$  is normally distributed among species with mean  $\mu$  and variance  $\sigma^2$ , whereas the mean number of individuals sampled from a species with abundance  $e^x$  is  $\lambda = ve^x$ , where  $v$  is a measure of the sampling intensity. Hence,  $\ln \lambda = \ln v + x$  is normally distributed with mean  $\ln v + \mu$  and variance  $\sigma^2$ . It follows that the number of individuals is Poisson lognormally distributed among the species in the community with parameters  $\ln v + \mu$  and  $\sigma^2$ . From a single sample from a community the parameters  $\ln v + \mu$  and  $\sigma^2$  can now be estimated by maximum likelihood as proposed by Bulmer (1974) (Methods). However, the parameter  $\mu$  in the underlying abundance distribution can not be estimated unless the sampling intensity is known or can be estimated in some way. We see that this formulation explains the well known effect observed by Preston (1948) that the lognormal distribution is just translated without changing form as the sampling intensity increases since actually  $\ln v + \mu$  increases with  $v$  and the variance parameter is constant.

Overdispersion in the sampling can be modeled by assuming that the distribution of  $N$  given  $\lambda$  is itself a Poisson lognormal distribution rather than a Poisson. Then,  $\lambda$  is first multiplied by a factor  $Z$  specific for each sampling, and  $N$  is assumed to be Poisson distributed with parameter  $\lambda Z$  for a given  $Z$ . If  $\ln Z$  is normally distributed with mean  $-\theta^2/2$  and variance  $\theta^2$ , then  $Z$  is lognormally distributed with mean 1. Now the parameter of the Poisson,  $\mu^x Z = e^{\ln v + x + \ln Z}$ , is lognormally distributed with parameters  $\ln v + \mu - \theta^2/2$  and  $\sigma^2 + \theta^2$ , implying that the unconditional distribution of  $N$  is the Poisson lognormal with parameters  $\ln v + \mu - \theta^2/2$  and  $\sigma^2 + \theta^2$ .

The parameter  $\theta$  is then a measure of over-dispersion in the sampling. More precisely, conditionally on the abundance, we have

$$\frac{\text{var}(N|x) - E(N|x)}{[E(N|x)]^2} = e^{\theta^2} - 1$$

If  $\theta = 0$  we have simple Poisson sampling.

### Bivariate Poisson lognormal model

These definitions can easily be generalized to two dimensions representing two communities. For given values of  $\lambda_1$  and  $\lambda_2$  we then assume that  $N_1$  and  $N_2$  are independent Poisson variates with mean values  $\lambda_1$  and  $\lambda_2$ . If the bivariate distribution of the mean values is  $f(\lambda_1, \lambda_2)$  the unconditional distribution is the two-dimensional Poisson mixture

$$P(N_1 = n_1, N_2 = n_2) = p_{n_1, n_2} = \iint P(N_1 = n_1, N_2 = n_2|\lambda_1, \lambda_2)f(\lambda_1, \lambda_2)d\lambda_1 d\lambda_2$$

If the log of the mean values  $\ln \lambda_1$  and  $\ln \lambda_2$  has the binormal distribution with marginal means and variances  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  and correlation  $\rho$ , the unconditional bivariate distribution of  $(N_1, N_2)$  is the bivariate Poisson lognormal distribution with parameters  $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho)$ . By the same kind of substitution as for the one-dimensional case, this distribution can be written as

$$p_{n_1, n_2} = \frac{e^{n_1 \mu_1 - e^{\mu_1}} + n_2 \mu_2 - e^{\mu_2}}{n_1! n_2!} \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{n_1 \sigma_1 u_1 - e^{-\sigma_1 u_1}} + n_2 \sigma_2 u_2 - e^{-\sigma_2 u_2} \times g(u_1, u_2; \rho) du_1 du_2$$

where  $g(u_1, u_2; \rho)$  is the binormal distribution with mean values zero, variances 1, and correlation  $\rho$ . As in the one-dimensional case we find that samples from two communities sampled with intensities  $v_1$  and  $v_2$  will follow the above bivariate Poisson lognormal distribution with  $\mu_1$  and  $\mu_2$  replaced by  $\ln v_1 + \mu_1$  and  $\ln v_2 + \mu_2$ .

In the two-dimensional case we see that the over-dispersion, defined as in the one-dimensional case, does not change the covariance between  $\ln \lambda_1$  and  $\ln \lambda_2$ , but their variances are increased by  $\theta^2$ . Hence, the correlation parameter in the two-dimensional Poisson lognormal will be reduced as expressed by Eq. 1.

## Methods

### Description of the data and sampling procedure

Beetles were sampled in altogether 12 different localities in southern Norway in the summers of 2004 (2 locations), 2005 (5 locations) and 2006 (5 locations). Selection of locality was based on inventories from the forestry sector and the municipalities. The sites were selected in order to cover a substantial part of the gradient from east to west in the distribution of oak in Norway; therefore the search for localities was limited to a few municipalities in each of the biogeographically relevant counties. Within the municipalities, criteria for selection included locations with at least five old and hollow oak trees. In each location, one hollow oak tree with minimum diameter 30 cm (1.3 m above ground) was selected randomly and the four hollow oaks

(also >30 cm diameter) closest to it was included. Exact location was recorded using GPS.

Each oak was sampled by means of two window traps (sensu Kaila 1993; windows measuring 20 × 40 cm), one mounted outside the opening of the tree hollow and the other hanging from branches in the canopy. The traps were operating from late May to early August and emptied monthly. This paper is based on those Red Listed (Kålås et al. 2006) and some additional rare beetle species that are associated with oak.

## Parameter estimation

As originally proposed by Fisher et al. (1943), and applied to the lognormal species abundance model by Bulmer (1974), the distribution of counts must be conditioned on the event that the species is observed. This is because the number of species with no representatives, and hence also the total number of species in the community, is usually unknown. Estimation of parameters based on one sample from a community is therefore done by fitting the zero-truncated Poisson lognormal distribution  $p_n/(1-p_0)$  by maximum likelihood. Hence, if there are  $R_i$  species with  $i$  representatives in the sample,  $i=1, 2, \dots$ , then the log likelihood takes the form

$$\ln L = \sum_i R_i \ln p_i - S \ln(1 - p_0)$$

where  $S = \sum R_i$  is the total number of species in the sample. This function must finally be maximized numerically with respect to the two unknown parameters in the univariate Poisson lognormal distribution.

If there are  $s$  species in the community the expected number of observed and non-observed species in the sample is  $ES = s(1 - p_0)$  and  $sp_0$ , respectively. The estimation equation for  $\hat{s}$  is accordingly  $S = \hat{s} (1 - \hat{p}_0)$ , known as Bulmer's method for estimating the number of species in the community (Bulmer 1974).

Estimation in the bivariate model is performed in the same way using the distribution conditioned on the species being observed in at least one of the samples, that is  $p_{i,j}/(1 - p_{0,0})$ . Writing  $R_{ij}$  for the number of species with  $i$  and  $j$  representatives in the first and second sample, respectively, the log likelihood is then

$$\ln L = \sum_{i,j} R_{ij} \ln p_{i,j} - S \ln(1 - p_{0,0})$$

where now  $S = \sum_{i,j} R_{i,j}$  is the number of different species observed in the two samples. This function can now be maximized numerically with respect to the five parameters of the bivariate Poisson lognormal distribution.

The parameters can be estimated by the R-package (R Development Core Team 2007) `poilog`.

## $\alpha$ -, $\beta$ - and $\gamma$ -diversity

Following Engen et al. (2002a) we compute estimates of the expected diversity within quadrats of different sizes. For small quadrats we then find the  $\alpha$ -diversity, whereas large quadrats with side-length around the largest distance between communities ( $\approx 300$  km) in the data set represents

the  $\gamma$ -diversity for that region. Approximating the distribution of abundance in quadrats by a log-normal distribution, the variance in the log-normal abundance distribution for the total community within a quadrat with side-length  $\alpha$  is (Engen et al. 2002b)

$$V(a) = \ln \int_0^{\sqrt{2}} e^{\sigma^2 \rho(az)} f(z) dz$$

where  $\rho(az)$  is the correlation between two communities at distance  $az$  corrected for over-dispersion and  $\sigma^2$  is the variance of the lognormal species abundance distribution at a single location. The function  $f(z)$  is the distribution of the distance between two randomly chosen points in a quadrat with side-length 1, which is (Engen et al. 2002b)

$$f(z) = \begin{cases} 2z(\pi + z^2 - 4z) & \text{for } 0 \leq z \leq 1 \\ 2z[2\arcsin(2/z^2 - 1) + 4(z^2 - 1)^{1/2} - 2 - z^2] & \text{for } 1 \leq z \leq \sqrt{2} \end{cases}$$

Using the variance  $V(a)$  and the lognormal approximation for the mean abundance in a quadrat one can simulate the species abundances within the quadrat, compute the relative abundances  $p_i$  for all species and the Shannon information index (Magurran 2004)  $H_{\text{sim}}(a) = -\sum p_i \ln p_i$ . The species abundances  $\lambda_i$  are then simulated from the lognormal distribution, first simulating the  $x_i = \ln \lambda_i$  as normal variates with zero mean and variance  $V(a)$ . The mean can be chosen arbitrarily because it only generates a constant factor in the abundances and hence does not affect relative abundances. This simulation can be repeated a large number of times (say 10 000), giving the mean value of  $H(a)$  as an estimate of the mean diversity for quadrats with side-length  $a$ . Using 300 km as our maximum distance we then have that the  $\alpha$ -diversity is  $H_\alpha = H(0)$ , the  $\gamma$ -diversity  $H_\gamma = H(300)$  and the  $\beta$ -diversity is the difference  $H_\beta = H_\gamma - H_\alpha$ .

## Results

### Total data set

Combining all local samples into one single community sample and fitting a zero-truncated Poisson lognormal distribution gives an estimated variance  $\hat{\sigma}^2 = 5.07$  of the lognormal with standard error 1.59 computed by parametric bootstrapping. Notice that this estimate also contains the over-dispersion  $\theta^2$  so that the true variance in log abundance is somewhat smaller (see section on over-dispersion in the sampling). The fitted abundance distribution is shown in Fig. 1, together with the data grouped in octaves. The octave  $R=1$  contains all species which are only observed once in the entire collection of beetles,  $R=2$  are those with 2–3 representative,  $R=3$  those with 4–7 representatives and so on. This model appears to fit well to the data. However, since there are as many as 41 species that are observed only once, the data reveal only a small fraction of the lognormal distribution and we cannot really say much about the fit for the large number of very rare species not included in the sample. The total number of observed species is 129. The maximum likelihood estimate of the zero term in the Poisson lognormal distribution is  $\hat{p}_0 = 0.519$  and the estimated total number of species from the defined sub-community is accordingly  $129/(1 - \hat{p}_0) = 268$

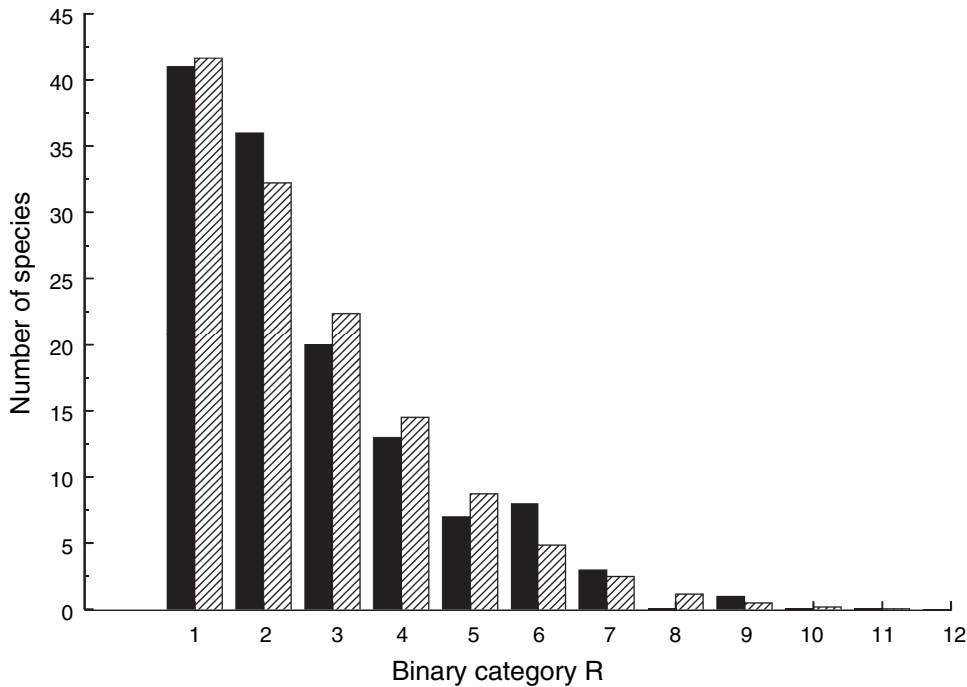


Fig. 1. Observed species abundance distribution (filled bars) grouped in octaves 1, 2–3, 4–7, ...  $2^{n-1} - (2^n - 1)$  for the entire collection of beetles together with the fitted Poisson lognormal abundance distribution multiplied with the estimated number of species which is 268. There are 129 observed species and estimated parameters of the underlying lognormal distribution are  $\hat{\sigma}^2 = 5.07$  and  $\hat{\mu} = -0.65$ .

with a standard error of 76 found by parametric bootstrapping.

Assuming approximately Poisson sampling of the rare species in the total accumulated sample, the total relative abundance of the species that are not observed can be estimated as the ratio between the number of singletons and the total number of individuals which is  $41/1499 = 0.0274$  (Good 1953, Engen 1975). This means that if the same kind of sampling process as the one giving our data set were repeated independently of the first sampling, then species previously not observed are altogether likely to have around 41 individuals in the sample. As these species would be rare, we should expect the number of additional species to be quite large by this doubling of the sampling intensity. Although the estimate of the total number of species in the community is rather uncertain, the estimate of the total abundance of those species is approximately unbiased and rather certain (Engen 1978).

### Estimation of variances

Figure 2 shows the variance of the lognormal estimated separately from each site, as well as the estimates obtained from fitting pair of sites, as function of the number of individuals in the sample for all samples with at least 20 individuals. We see that there is only a small increase in uncertainties with decreasing sample size and that the uncertainty for the smallest sample sizes of about 20 individuals still is surprisingly small. One reason for the small sensitivity of uncertainties with respect to number of individuals sampled is that the number of observations from the Poisson lognormal distribution is the number of species and not the number of individuals. Because the observed

species number only increases rather slowly with increasing sample size (Fig. 3), uncertainties in the estimated variances do not vary much with observed individual number. The estimates distribute themselves well with no large bias around a common value. The mean of these estimates was 4.44 (SD = 0.11) which is smaller than the variance found from the total accumulation of all individuals sampled. The mean of the variance estimates obtained from pairs of communities is 4.26 (SD = 0.09), not significantly different from the above mean estimate for each site, in accordance with a hypothesis of practically no spatial variation in community parameters.

A spatial segregation of species can also be visualized by calculating the expected number of species as function of the expected number of individuals in subsamples (so-called rarefaction curve, sensu Colwell and Coddington 1994 and Ugland et al. 2003) from the total community. In practice, this is done by varying the mean value parameter  $\ln v + \mu$  in the Poisson lognormal model fitted to the total data set keeping the variance parameter constant equal to the estimate. In Fig. 3 we show this curve together with the observed number of species against observed individual number for all sites. The observations tend to show somewhat smaller species number than predicted by the rarefaction curve, but only for the largest samples, which is also an indication that communities are only to a minor degree spatially segregated.

### Analysis of similarity of samples

Figure 4 shows the estimated correlations between all pairs of communities against the distance between them. The curves shown in the graph are fitted by minimizing sum of

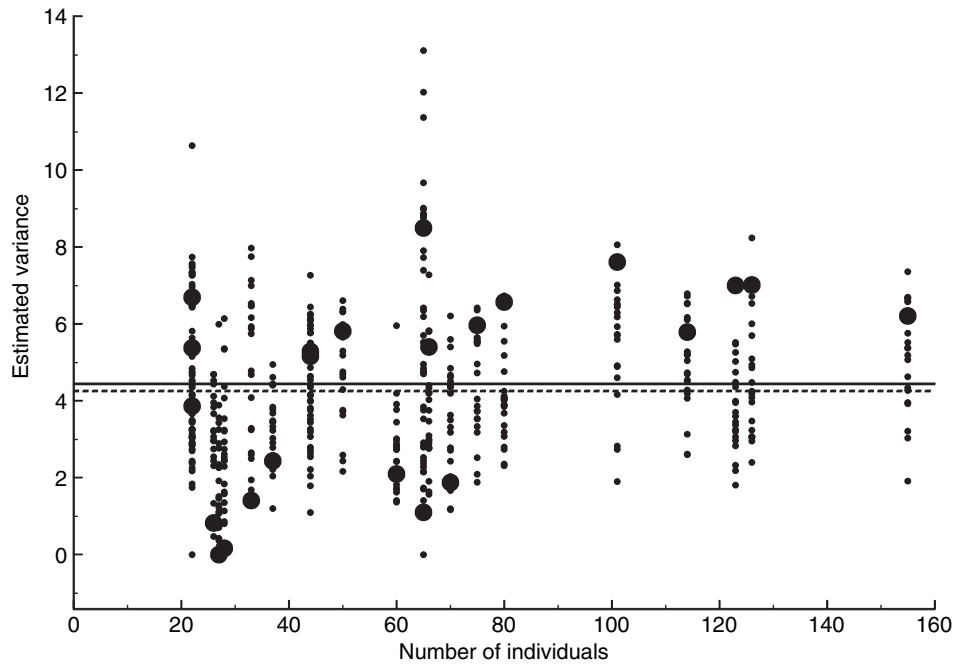


Fig. 2. Estimates of the variance of the lognormal from fitting the bivariate model pairs of communities (small dots) with mean value 4.26 (dotted line), and from the univariate model to each single community (large dots) with mean value 4.45 (solid line).

squares. The correlation at zero distance is estimated to be 0.853 (SD=0.036) by fitting the Gaussian type of correlation function which fits slightly better than the exponential function. From the relation between  $\rho$  and  $\rho_{xy}$  we estimate the over-dispersion  $\theta^2$  at local sites to be 0.65 (SD=0.19), which is a rather large over-dispersion. This measure of over-dispersion includes sampling of individuals as well as the effect of local density variation, that is,

variation in individual numbers among hollow oak trees at a given location around a mean value for the local area. Together, these two effects make the number of individuals sampled for a given species showing over-dispersion defined by  $\theta^2$  relative to Poisson distribution. Subtracting the over-dispersion component from the estimated variance obtained as the mean over all locations, we find that the variance parameter in the underlying lognormal species

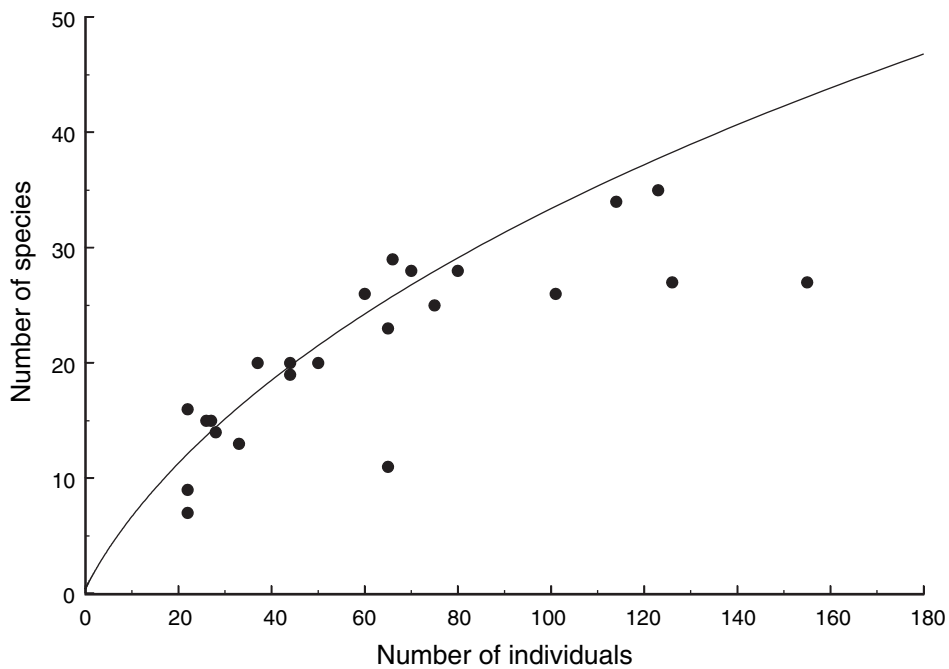


Fig. 3. Observed number of species plotted against observed number of individuals for each sample. The solid line is the estimated mean number of species observed as function of individual number under the assumption that the samples are random samples from the total community (rarefaction curve).

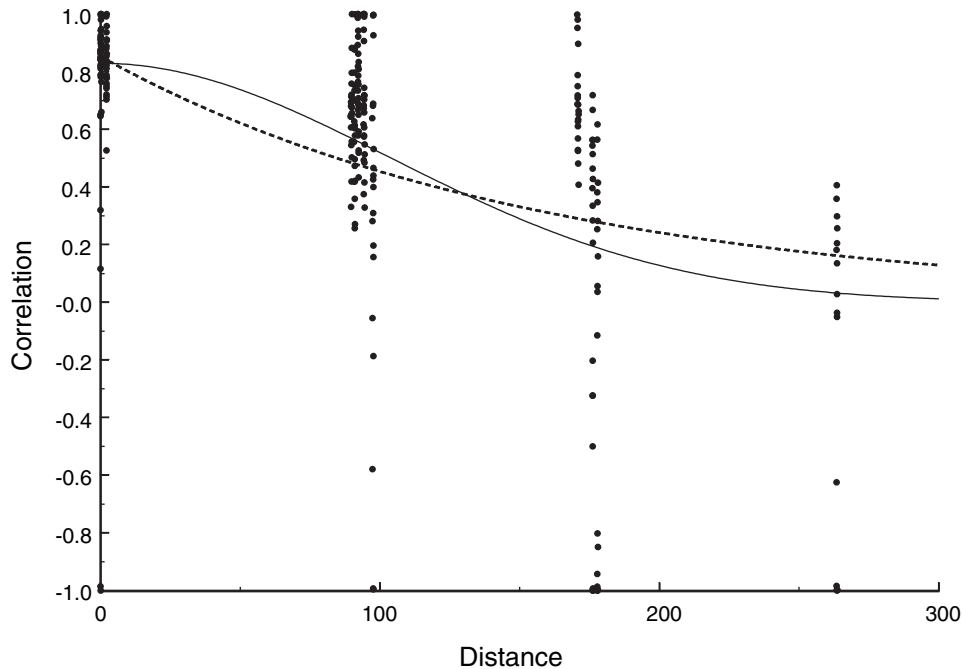


Fig. 4. Estimated within years similarity between communities, measured as the correlation parameter in the bivariate Poisson lognormal distributions, plotted against distance in kilometers. The solid and dashed line are the Gaussian and exponential spatial correlation function respectively, both fitted by least square.

abundance at each single site is 3.79. We have no information about the over-dispersion for the total dataset, but this is likely to be quite small since the values of the sampling variable  $Z$  for a given species vary from sample to sample so that the total variation will be small from the law of large numbers. The large over-dispersion indicates that communities at very close distance sampled the same season will be rather different. This can only mean that the local density of a species shows large variation even over short distances. Thus, the number of individuals of the same species in almost parallel samples will vary a lot.

The correlations drop rather quickly with distance between the samples with a spatial scaling of order 200 km. At distance 300 km the correlations seems to distribute themselves around zero (Fig. 4). Hence, there is some spatial structure in the data in the sense that quite different species may be the most abundant at two locations at distances larger than around 200 km. This again indicates large spatial patchiness of single species most likely due to spatial heterogeneity in the habitat with a scale larger than the one shown by this experiment. However, we have seen above that the structure of the community of this selected set of rare species seems rather constant in space.

The sampling effect of the correlation between locations is demonstrated in Fig. 5, in which the probability that a species should be found in a sample is calculated conditioned on the number of representatives the species has in another sample. We see that if a species has one individual in a sample, the probability that it should be found in another nearby sample with correlation around 0.7 is as small as 0.25. And even if it has five representatives in the first sample, the probability that it will be represented in the other one is only about one half.

Engen et al. (2002a) and Lande et al. (2003) also proposed fitting the univariate Poisson lognormal to the total set of recordings of individual counts for each species and each location. If there is no variation in the sampling effort and no permanent spatial variation in the habitat giving variation in local carrying capacities (Engen et al. 2002c), the variance obtained by this method should be the same as the one obtained from each single site. Permanent spatial variation will lead to a large estimate. For the present data (Fig. 6), this variance is 3.885 (SD = 0.73), which is actually smaller, but not significantly smaller, than the mean estimate over all sites. Hence, it does not seem to be any significant permanent variation in the overall habitat quality.

Figure 7 shows the mean information index within quadrats of increasing size. The mean local  $\alpha$ -diversity is estimated to be  $H_\alpha = 3.90$ , while the  $\gamma$ -diversity for a quadrat with side-length 300 km, corresponding approximately to the largest distance in the data, is  $H_\gamma = 4,56$ , giving a  $\beta$ -diversity over areas of these size  $H_\beta = 0.66$ . The model indicates that the  $\gamma$ - and  $\beta$ -diversity will continue increasing by a further increase in area (Fig. 7, dotted line).

## Discussion

In this paper we use the bivariate lognormal species abundance model to estimate similarities of communities of oak-living beetles in southern Norway. We estimate the total relative abundance of species not observed to be about 2.7%, suggesting that the number of unrecorded species is quite large as many species are quite rare in the sample. Accordingly, analyses by the method of Bulmer (1974) based on the assumption of a lognormal species abundance

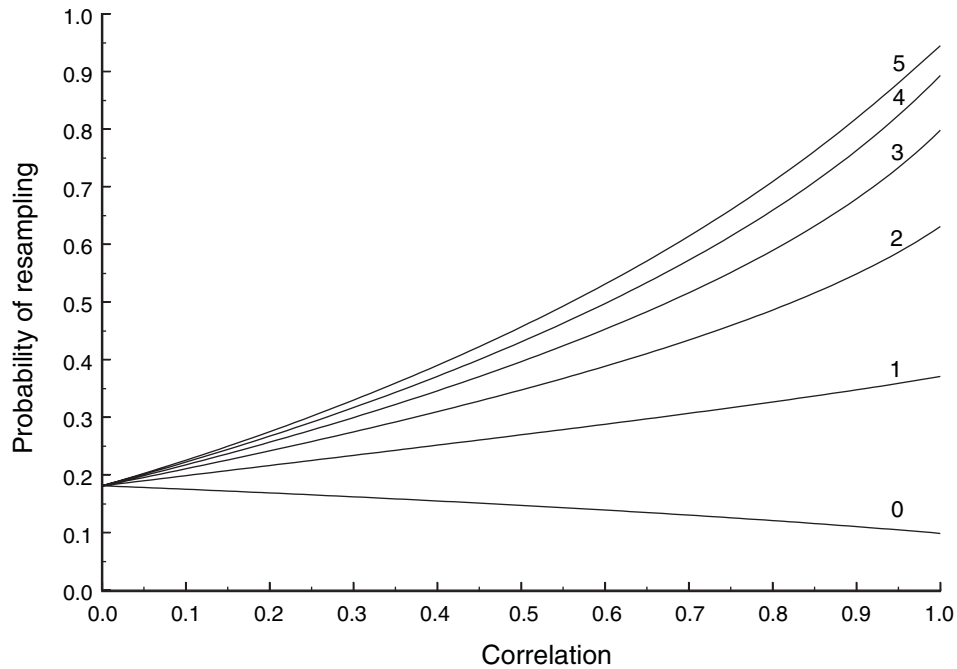


Fig. 5. Each line shows the probability that a species is present in a sample from a community conditioned on the number of representatives (0, 1, . . . , 5) it has in a sample from another community, as function of the correlation between the two communities.

distribution (Fig. 1) indicates that we recorded just under half (48.1%) of the species that actually are present in our sub-community.

The correlation between communities drops to small values at distances larger than about 200 km (Fig. 4). Furthermore, this correlation approaches a value smaller than 1 when the distances approaches zero, indicating that

there is over-dispersion in the sampling relative to the Poisson distribution. This over-dispersion is corrected for in the diversity calculations shown in Fig. 7. Our estimate of  $\theta^2$  of 0.65 shows that the over-dispersion is rather large, accounting for 15% of the estimated variance parameter of the Poisson lognormal model. If this sampling effect were not taken into account, the estimated variance of the log-

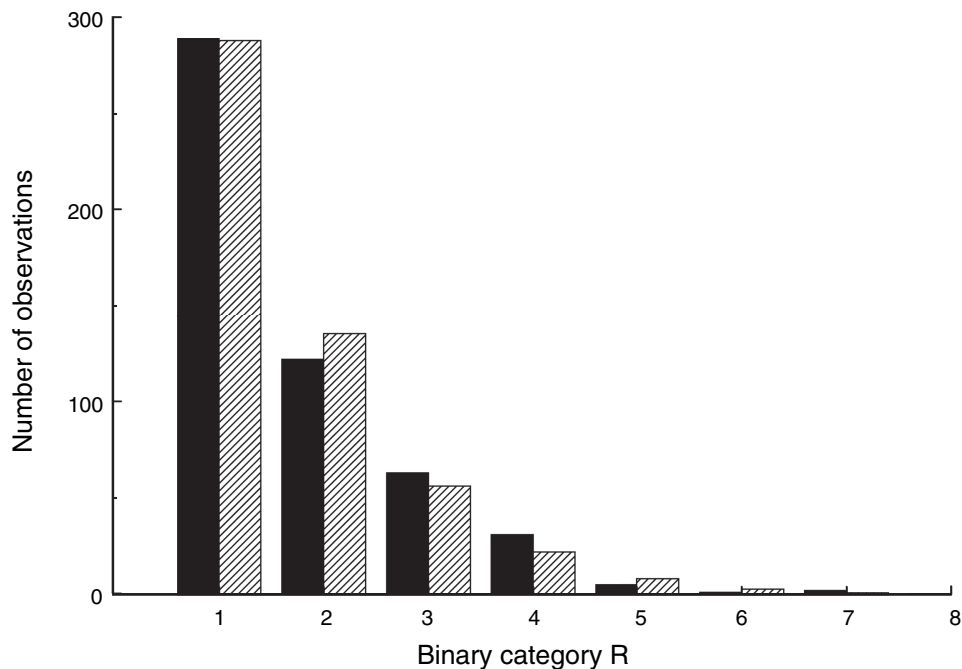


Fig. 6. The univariate Poisson lognormal model fitted to the total set of recordings of individual counts for each species at each location, that is, each individual count for each species at each location is considered an observation from the same Poisson lognormal distribution. The estimated variance of this lognormal is 3.89. The grouping into octaves is the same as in Fig. 1.

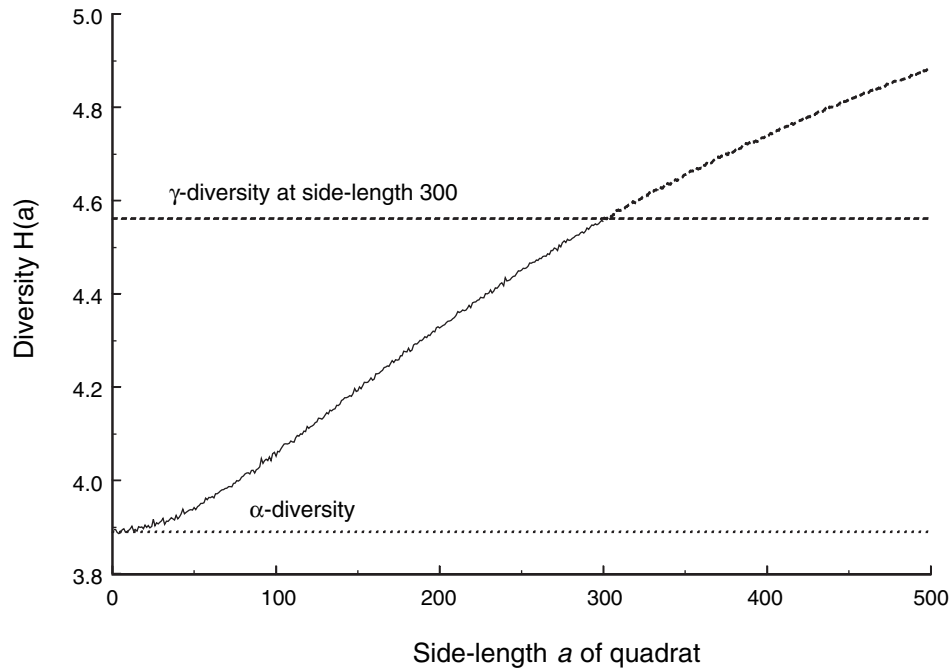


Fig. 7. Estimates of the information index of diversity for communities within quadrats of varying size. Small quadrats give the local  $\alpha$ -diversity. The solid line show the estimated  $\gamma$ -diversity for quadrat size up to a side-length of 300 km which is about the largest distance in the data. The dashed line shows the extrapolation of this curve to larger areas. The variance parameter used in the calculations is corrected for over-dispersion in the sampling relative to a Poisson distribution.

normal species abundance distribution would have been over-estimated by about 15%, and the diversities shown in Fig. 7 would be under-estimated as the diversity decrease with increasing variance (Bulmer 1974).

Our analysis is based on fitting the bivariate lognormal species abundance model for each pair of communities (Engen et al. 2002a) and using the estimated correlation, after correcting for over-dispersion due to sampling effects, as a measurement of similarity between communities. Similarity measures most commonly used in community ecology, such as the Jaccard and Sørensen index (Magurran 2004), has the disadvantage that they are often strongly affected by sampling intensity. Even the nice method of Chao et al. (2005) for bias correction gives some bias for random Poisson sampling. For over-dispersed sampling of the magnitude shown in the present study, the robustness against changing sampling intensities is not known. Although the lognormal model may be only approximately correct (Williamson and Gaston 2005), there are great advantages in using this parametric model, enabling elimination of sampling intensities and over-dispersion in sampling as well as application of the maximum likelihood procedure. The correlation also has a number of interesting relations to spatio-dynamical models for single species (Engen et al. 2002c) as well as communities (Engen et al. 2002a) and it clearly expresses the type of similarity measured by the more commonly used indices. One disadvantage is, however, that the computations are rather time consuming as each calculation of the likelihood is based on a large number of two-dimensional numerical integrations.

The lognormal distribution is often found to describe species abundance distributions of communities from a

large variety of taxa quite well (Preston 1948, 1962, 1980, Tokeshi 1993, McGill 2003, McGill et al. 2007). In our example of oak-living beetles this assumption seems justified (Fig. 1, 2, 6). However, assessing the correct distribution from species-abundance data can be difficult, e.g. due to the influence of the sampling process, resulting in low power of tests for discrimination between different models that give quite similar distributions (Etienne and Olf 2005, Williamson and Gaston 2005). This is probably best illustrated by the problems encountered in analyses of the validity of Hubbell's (2001) neutral theory when based on species distributions (McGill et al. 2006, Etienne 2007). It is, however, important to notice that our choice of a lognormal distribution is also based on analyses of stochastic dynamic models of temporal variation in community structure (Engen and Lande 1996). Here species enter the community according to a Poisson process and have dynamics of log abundances that can be described by the Ornstein-Uhlenbeck process (Karlin and Taylor 1981). At least in many vertebrates, such a loglinear model of density regulation is often appropriate (Sæther and Engen 2002, Sæther et al. 2007), and it has also been used to model temporal variation of insect populations (Royama 1992). As shown by Engen and Lande (1996) and Diserud and Engen (2000), this will generate a lognormal species abundance distribution. Further, Engen and Lande (1996) showed that even a heterogeneous model of this type, in which growth rates of species entering the community were normally distributed among species, resulted in a lognormal species abundance distribution.

Many studies of species diversity have focused on estimates of the total number of species that is present in an area using both non-parametric and parametric methods

(McGill et al. 2007). However, estimation of species richness in natural communities seems notoriously difficult (Palmer 1990, Gotelli and Colwell 2001), and may result in severely biased estimates (Lande 1996, Uglund et al. 2003, Walther and Moore 2005). Our approach based on the assumption of a lognormal species abundance distribution provide an approximately unbiased estimate of the total number of species if the lognormal model fits the community. However, this estimate (268 species) is rather uncertain due to the uncertainty in the shape of the non-observed left tail of the abundance distribution (Gray et al. 2005), but it is definitely an indication that there is quite a large number of species not yet observed during the experiment. Even with the assumption that the lognormal model fits all the way down to very small non-observed abundances, the standard error of the estimator found by parametric bootstrapping is as large as 76, so the real number of species may be substantially smaller, and also much larger, than the estimated number of 268 species.

The beetle data used in this study was collected as part of the Norwegian government-initiated 'National Program for Surveying and Monitoring Biodiversity'. The aim with the survey was to assess the occurrence of Red Listed species as well as identifying areas of particular occurrence of such species, so-called hot-spot localities. The present results clearly demonstrate many of the problems in assessing species diversity by such a sampling scheme that involves only one season of sampling in each locality. For instance, a large number of species was present in the sample by only one individual (Fig. 1). A high ratio of singletons is commonly found when sampling insect communities (Novotny and Basset 2000, Sverdrup-Thygeson and Ims 2002, Lindhe and Lindelow 2004). This means that the probability of recording a species by repeating the same sampling process is quite small, especially when there is over-dispersion in the sampling, as demonstrated in Fig. 5. Furthermore, the number of species will also increase with the number of sampled individuals (Fig. 3), as is generally the case (Pielou 1975, Lande 1996). As a consequence, identifying critical areas for conservation and the occurrence of rare species can be difficult using a sampling that is only based on trappings from a single or a few seasons. Results from extensive studies in boreal forest in Finland also show that rare and threatened beetle species accumulate slowly in the samples, and that large samples (>100 000 beetle individuals) may be necessary in order to discriminate between different areas concerning occurrence of Red Listed species (Martikainen and Kouki 2003, Martikainen and Kaila 2004).

As a response to these challenges in estimating species diversity, we suggest that a spatially extensive sampling should be combined by collection of data on variation in community structure over time. By using the theoretical framework by Lande and Engen (1996), Engen et al. (2002a) and Lande et al. (2003) such data will allow estimation of the heterogeneity among species (Engen 2007a, 2007b), that is, the parameters describing the dynamics may differ among species. We can then determine whether the relative abundance of a species is generally similar over time and in space, which will greatly facilitate estimation of many important community characteristics

such as turnover rates of species in space and time (Engen et al. 2002a, 2007c, Lande et al. 2003, Walla et al. 2004).

*Acknowledgements* – The R package 'poilog' is available at The Comprehensive R Archive Network <<http://cran.r-project.org/src/contrib/Descriptions/poilog.html>> or by just typing "install.packages('poilog')". This study was financed by a grant from the Directorate for Nature Management in Norway, by a grant from the Research Council of Norway (project 159571/V40) and the core funding from NTNU to the Centre for Conservation Biology (CCB).

## References

- Anderson, M. J. et al. 2006. Multivariate dispersion as a measure of beta diversity. – *Ecol. Lett.* 9: 683–693.
- Baltanás, A. 1992. On the use of some methods for the estimation of species richness. – *Oikos* 65: 484–492.
- Bulmer, M. G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. – *Biometrics* 30: 101–110.
- Chao, A. et al. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. – *Ecol. Lett.* 8: 148–159.
- Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity by extrapolation. – *Philos. Trans. R. Soc. Lond. B* 345: 101–118.
- Condit, R. et al. 2002. Beta-diversity in tropical forest trees. – *Science* 295: 666–669.
- Crist, T. O. and Veech, J. A. 2006. Additive partitioning of rarefaction curves and species-area relationships: unifying  $\alpha$ -,  $\beta$ - and  $\gamma$ -diversity with sample size and habitat area. – *Ecol. Lett.* 9: 923–932.
- Davidar, P. et al. 2007. The effect of climatic gradients, topographic variation and species traits on the beta diversity of rain forest trees. – *Global Ecol. Biogeogr.* 16: 510–518.
- Diserud, O. H. and Engen, S. 2000. A general and dynamic species abundance model, embracing the lognormal and the gamma models. – *Am. Nat.* 155: 497–511.
- Engen, S. 1975. The coverage of a random sample from a biological community. – *Biometrics* 31: 201–208.
- Engen, S. 1978. Stochastic abundance models. – Chapman and Hall.
- Engen, S. 2007a. Heterogeneous communities with lognormal species abundance distribution: species–area curves and sustainability. – *J. Theor. Biol.* 249: 791–803.
- Engen, S. 2007b. Heterogeneity in dynamic species abundance models: the selective effect of extinction processes. – *Math. Biosci.* 210: 490–507.
- Engen, S. and Lande, R. 1996. Population dynamic models generating the lognormal species abundance distribution. – *Math. Biosci.* 132: 169–183.
- Engen, S. et al. 2002a. Analysing spatial structure of communities by the two-dimensional Poisson lognormal species abundance model. – *Am. Nat.* 160: 60–73.
- Engen, S. et al. 2002b. The spatial scale of population fluctuations and quasi-extinction. – *Am. Nat.* 160: 439–451.
- Engen, S. et al. 2002c. Migration and spatiotemporal variation in population dynamics in a heterogeneous environment. – *Ecology* 83: 570–579.
- Etienne, R. S. 2007. A neutral sampling formula for multiple samples and an 'exact' test of neutrality. – *Ecol. Lett.* 10: 608–618.
- Etienne, R. S. and Olff, H. 2007. Confronting different models of community structure to species-abundance data: a Bayesian model comparison. – *Ecol. Lett.* 8: 493–504.

- Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? – *Syst. Biol.* 51: 331–363.
- Fisher, R. A. et al. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. – *J. Anim. Ecol.* 12: 45–58.
- Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. – *Biometrika* 40: 237–264.
- Golicher, D. J. et al. 2006. Lifting a veil on diversity: a Bayesian approach to fitting relative-abundance models. – *Ecol. Appl.* 16: 202–212.
- Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – *Ecol. Lett.* 4: 379–391.
- Gray, J. S. et al. 2005. The impact of rare species on natural assemblages. – *J. Anim. Ecol.* 74: 1131–1139.
- Grundy, R. M. 1951. The expected frequencies in a sample of an animal population in which the abundances are lognormally distributed. – *Biometrika* 38: 427–434.
- Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.
- Kålås, J. A. et al. 2006. 2006 Norwegian Red List. – Artsdata-banken, Norway.
- Kaila, L. 1993. A new method for collecting quantitative samples of insects associated with decaying wood or wood fungi. – *Entomol. Fenn.* 4: 21–23.
- Karlin, S. and Taylor, H. M. 1981. A second course in stochastic processes. – Academic Press.
- Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – *Oikos* 76: 5–13.
- Lande, R. et al. 2003. Stochastic population dynamics in ecology and conservation. – Oxford Univ. Press.
- Legendre, P. et al. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. – *Ecol. Monogr.* 75: 435–450.
- Levins, R. 1968. Evolution in changing environments. – Princeton Univ. Press.
- Lindhe, A. and Lindelow, A. 2004. Cut high stumps of spruce, birch, aspen and oak as breeding substrates for saproxylic beetles. – *For. Ecol. Manage.* 203: 1–20.
- Ludovisi, A. and Taticchi, M. I. 2006. Investigating beta diversity by Kullback-Leibler information measures. – *Ecol. Modell.* 192: 299–313.
- Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.
- Mao, C. X. and Colwell, R. K. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. – *Ecology* 86: 1143–1153.
- Martikainen, P. and Kouki, J. 2003. Sampling the rarest: threatened beetles in boreal forest biodiversity inventories. – *Biodiv. Conserv.* 12: 1815–1831.
- Martikainen, P. and Kaila, L. 2004. Sampling saproxylic beetles: lessons from a 10-year monitoring study. – *Biol. Conserv.* 120: 171–181.
- McGill, B. J. 2003. A test of the unified neutral theory of biodiversity. – *Nature* 422: 881–885.
- McGill, B. J. et al. 2006. Empirical evaluation of neutral theory. – *Ecology* 87: 1411–1423.
- McGill, B. J. et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. – *Ecol. Lett.* 10: 995–1015.
- Melis, C. et al. 2007. The role of moose *Alces alces* L. in boreal forest—the effect on ground beetles (Coleoptera, Carabidae) abundance and diversity. – *Biodiv. Conserv.* 16: 1321–1335.
- Novotny, V. and Basset, Y. 2000. Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. – *Oikos* 89: 564–572.
- Novotny, V. et al. 2007. Low beta diversity of herbivorous insects in tropical forests. – *Nature* 448: 692–697.
- O’Hara, R. B. 2005. Species richness estimators: how many species can dance on the head of a pin? – *J. Anim. Ecol.* 74: 375–386.
- Palmer, M. W. 1990. The estimation of species richness by extrapolation. – *Ecology* 71: 1195–1198.
- Pielou, E. C. 1975. Ecological diversity. – Wiley.
- Preston, F. W. 1948. The commonness, and rarity, of species. – *Ecology* 29: 254–283.
- Preston, F. W. 1962. The canonical distribution of commonness and rarity. – *Ecology* 43: 185–215, 410–432.
- Preston, F. W. 1980. Noncanonical distributions of commonness and rarity. – *Ecology* 61: 88–97.
- Qian, H. and Ricklefs, R. E. 2007. A latitudinal gradient in large-scale beta diversity for vascular plants in North America. – *Ecol. Lett.* 10: 737–744.
- R Development Core Team 2007. R: A language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <<http://www.R-project.org>>
- Rooney, T. P. et al. 2007. Biotic homogenization and conservation prioritization. – *Biol. Conserv.* 134: 447–450.
- Royama, T. 1992. Analytical population dynamics. – Chapman and Hall.
- Sæther, B.-E. and Engen, S. 2002. Pattern of variation in avian population growth rates. – *Philos. Trans. R. Soc. B* 357: 1185–1195.
- Sæther, B.-E. et al. 2007. Predicting fluctuations of re-introduced ibex populations: the importance of density-dependence, environmental stochasticity and uncertain population estimates. – *J. Anim. Ecol.* 76: 326–336.
- Sverdrup-Thygeson, A. and Ims, R. A. 2002. The effect of forest clearcutting in Norway on the community of saproxylic beetles on aspen. – *Biol. Conserv.* 106: 347–357.
- Tokeshi, M. 1993. Species abundance patterns and community structure. – *Adv. Ecol. Res.* 24: 111–186.
- Ugland, K. I. et al. 2003. The species-accumulation curve and estimation of species richness. – *J. Anim. Ecol.* 72: 888–897.
- Walla, T. R. et al. 2004. Modeling vertical beta-diversity in tropical butterfly communities. – *Oikos* 107: 610–618.
- Walther, B. A. and Moore, J. L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. – *Ecography* 28: 815–829.
- Whittaker, R. H. 1970. Communities and ecosystems. – Macmillan.
- Whittaker, R. H. 1972. Evolution and measurement of species diversity. – *Taxon* 21: 213–251.
- Williamson, M. and Gaston, K. J. 2005. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. – *J. Anim. Ecol.* 74: 409–422.