



## Estimating similarity of communities: a parametric approach to spatio-temporal analysis of species diversity

Steinar Engen, Vidar Grøtan and Bernt-Erik Sæther

*S. Engen (steinaen@math.ntnu.no), Centre for Conservation Biology, Dept of Mathematical Sciences, Norwegian Univ. of Science and Technology, NO-7491 Trondheim, Norway. – V. Grøtan and B.-E. Sæther, Centre for Conservation Biology, Dept of Biology, Norwegian Univ. of Science and Technology, NO-7491 Trondheim, Norway.*

Several stochastic models with environmental noise generate spatio-temporal Gaussian fields of log densities for the species in a community. Combinations of such models for many species often lead to lognormal species abundance distributions. In spatio-temporal analysis it is often realistic to assume that the same species are expected to occur at different times and/or locations because extinctions are rare events. Spatial and temporal  $\beta$ -diversity can then be analyzed by studying pairs of communities at different times or locations defined by a bivariate lognormal species abundance model in which a single correlation occurs. This correlation, which is a measure of similarity between two communities, can be estimated from samples even if the sampling intensities vary and are unknown, using the bivariate Poisson lognormal distribution. The estimators are approximately unbiased, although each specific correlation may be rather uncertain when the sampling effort is low with only a small fraction of the species represented in the samples. An important characteristic of this community correlation is that it relates to the classical Jaccard- or the Sørensen-indices of similarity based on the number of species present or absent in two communities. However, these indices calculated from samples of species in a community do not necessarily reflect similarity of the communities because the observed number of species depends strongly on the sampling intensities. Thus, we propose that our community correlation should be considered as an alternative to these indices when comparing similarity of communities. We illustrate the application of the correlation method by computing the similarity between temperate bird communities.

Species diversity is an emergent property of communities that is of crucial importance for conservation of biodiversity as well as for understanding of trophic interactions and ecosystem processes. This has led to much work in developing indices of similarity that can be used to compare differences between communities or samples from communities (Peet 1974, Pielou 1975, Koleff et al. 2003, Magurran 2004). The first attempts leading to the classical Jaccard (1912)- and Sørensen (1948)-indices were both based on the number of species present in both samples and the numbers only seen in each of them. Although presence and absence should refer to the total community, these indices are most commonly used by counting species present in samples.

The Jaccard- and Sørensen-indices relate directly to Whittaker's (1970, 1972) partitioning of species diversity into the  $\alpha$ -diversity component within localities,  $\gamma$ -diversity in the whole region and  $\beta$ -diversity defined as the turnover of species among samples at different localities generating the difference between the local and regional diversity. A large number of studies have used this decomposition to estimate relative contribution of these diversity components (see references in Koleff et al. [2003] and McGill et al.

[2007]). Whittaker's decomposition is closely related to the concept of similarity as large  $\beta$ -diversity implies that samples will show decreasing similarity when recorded at increasing spatial distance (Engen et al. 2008).

In a pioneering study, Fisher et al. (1943) presented the idea of describing communities by studying the distribution of species abundances among the species in a community. Although Fisher initiated his analysis by assuming that abundances were gamma distributed and thus described by two parameters only, he showed surprisingly that even the simpler one-parameter model obtained as the shape parameter approached zero fitted well to a large number of communities. Preston (1948, 1960, 1962, 1980) considered an alternative parametric approach by empirically showing that the log-normal species abundance model fitted well to a large number of species abundance distributions covering communities from a wide range of taxa. Taking into account how complex the spatio-temporal dynamics of communities with a large number of species is likely to be, the most interesting findings of these pioneers is the simplicity of these purely descriptive models and the fact that the same kind of patterns seems to be applicable to a large number of different communities, suggesting a

relatively general description of the distribution of species abundances (Tokeshi 1993).

More recent research on the structures of communities has focused on underlying population dynamical mechanisms, possibly including speciations and extinctions, generating patterns of similar types as those described by the above pioneers. Such spatio-temporal models are naturally classified into two major categories. One is the group of neutral models using essentially Fisher's model to describe the species abundance distribution in a large meta-community together with an analysis of local community dynamics by migration and local extinctions (Hubbell 2001). Local dynamics is described by demographic stochasticity only, that is, birth and deaths of individuals are independent events among individuals. Furthermore, the neutrality assumption also implies that the dynamics of all species are described by the same mathematical rules so that all species and individuals are actually interchangeable in the models (Leigh 2007). The other category contains models where the dynamics of species abundances are driven by environmental stochasticity generated by a fluctuating environment affecting all individuals of a species in the same or a similar way. Engen and Lande (1996a) presented a model of this type with speciation and extinction and linear dynamics on the log scale, leading to a lognormal species abundance distribution. In a parallel paper Engen and Lande (1996b) also showed that a similar kind of model with somewhat different assumptions for the density regulation generated the gamma distribution including the special case of Fisher's model as well as the extended form of the gamma model with negative shape parameter in the interval  $(-1, 0)$  (Engen 1974, 1978). Heterogeneity, that is, parameters describing dynamics differing among species, was also included in the lognormal model of Engen and Lande (1996a). Later this class of models has been extended to describe spatial as well as temporal species abundance fluctuations (Engen et al. 2002, 2008, Lande et al. 2003, Engen 2007a, b), models with environmental as well as demographic stochasticity (Loreau and de Mazancourt 2008), and models including competition between species (Mutshinda et al. 2008, 2009). Engen (2001) showed that even dispersal, described as smooth continuous diffusion of individuals, could be included in linear models on the log scale in such a way that log abundances of a single species was still a Gaussian spatial field and pointed out that such models with migration can also generate lognormal spatio-temporal species abundance models. An important feature of all these spatio-temporal lognormal species abundance models is that the community autocorrelation can be used to partition species diversity following Whittaker (1970, 1972) into  $\alpha$ -,  $\beta$ - and  $\gamma$ -components (Engen et al. 2002, 2008, Lande et al. 2003).

Many of the above models with environmental stochasticity generate communities in which the log abundance of single species are Gaussian fields in space and time. Further, homogeneous as well as classes of heterogeneous models lead to species abundances being lognormally distributed among species at a given location and time. Considering two different locations possibly at different times, the distribution of log abundances among species is then the bivariate normal distribution defined by 5 parameters, the means and variances as well as the correlation. Engen

et al. (2002, 2008), Lande et al. (2003) and Walla et al. (2004) utilized this class of models to perform temporal and spatial analysis of communities by studying the correlation structure obtained by considering the correlations between all pairs of observed communities in space and time. Even if noise components that affect all species in the same way will make the abundances of species dependent, Engen and Lande (1996a) showed that such noise is confounded with the mean values and will not affect the correlations. Assuming that the abundances are independent among species therefore gives a correct estimate of the correlation even if such common noise components are operating (for details see Engen and Lande [1996a] and Lande et al. [2003]). The realism of assuming independent dynamics apart from the common noise terms is also supported by Mutshinda et al. (2009), who estimated these correlations to be quite small.

The lognormal species abundance model has been widely applied in analyses of how structures of communities vary in space and time (May 1975, Ugland and Gray 1982, Tokeshi 1993, Engen et al. 2002, 2008, Lande et al. 2003, Magurran 2004, McGill et al. 2007). Here we focus on the analysis of spatial and temporal structure of communities when any two communities can be described by a bivariate log-normal species abundance distribution, that is, the log-abundances  $(X, Y)$  in the two communities follow a bivariate normal distribution among species. As the marginal distributions are normal, this is in agreement with Preston's (1948, 1960) lognormal species abundance model when each community is considered separately as well as the Poisson lognormal model for samples based on Poisson sampling (Bulmer 1974, O'Hara 2005, Golicher et al. 2006).

The community correlation described above is a measure of similarity between two communities. If the correlation equals one, the communities are identical. If it is zero, they are independent in structure. A negative correlation expresses that species that are abundant in one community are likely to be rare in the other. However, this measurement is based on the assumption that the species at the two locations are the same, so the Jaccard- and the Sørensen-indices for the total communities as well as the abundance based analogy of these indices defined by Chao et al. (2005), are equal to 1. This is a realistic assumption as long as the communities are connected spatially by migration or when a given community with temporal fluctuations in species abundance are studied through time. There is no lower limit for the log abundance in the model, so a species present in one community sample may have such a small abundance in the other that it is practically impossible to observe even under intense sampling. Hence, even large samples may have species common to both communities as well as species present in only one of them so that the Jaccard- and the Sørensen-indices estimated directly from the samples can be substantially smaller than 1. For very diverse communities, the Jaccard- and the Sørensen-indices tend to be very close to 1 when the sample sizes get extremely large, so the assumption of a common set of species including many species with abundance that is practically zero, but still positive, is often likely to be a realistic model (Magurran 2007). In particular, when studying temporal fluctuations in community composition based on similarities estimated from time series of

community samples, this assumption seems realistic since extinctions are rare events.

In this paper we present the correlation in log abundances between pairs of communities as a simple index of similarity. We show how this correlation can be estimated by maximum likelihood from samples of communities using free available software. We then show how the community correlation, based on the assumption of a Gaussian field of log densities for the species in communities, relates to other measures of community similarity (Koleff et al. 2003, Magurran 2004, Chao et al. 2006). Finally, we illustrate the practical applicability of our approach by estimating the community correlation in two studies of bird communities.

## Estimation

Fisher et al. (1943) defined species abundances as continuous variables that could take any positive value. As they mainly analyzed individuals caught in light traps the abundance of a species was defined as the mean number of individuals caught by some given sampling effort which is a measurement of the density of individuals. For a given species with a given abundance Fisher et al. assumed that the number of individuals observed was Poisson distributed with mean value proportional to the abundance of the species. By this assumption of abundances being gamma distributed the number of individuals caught then has a negative binomial distribution among species. Because the number of species that are not observed is unknown, Fisher et al. used the zero-truncated distribution for which he derived the limiting distribution as the shape parameter was sent to zero and the number of species in the community approached infinity, obtaining what is known as Fisher's log series distribution.

Using Fisher et al.'s approach applied to two communities, each species is represented by a realization of a bivariate random variable expressing its abundance in the two communities. The natural generalization of the lognormal species abundance model to two communities is accordingly the bivariate lognormal distribution. Assuming Poisson sampling, as proposed by Fisher et al. (1943), estimation of the correlation  $\rho$  can be performed by maximum likelihood using the bivariate Poisson lognormal distribution (Engen et al. 2002, Lande et al. 2003, Walla et al. 2004) truncated to exclude the double-zero observations. Let us first consider a sample from one community described by a distribution of densities among species at a given location and time. Assuming random sampling the number of individuals sampled of a species with log abundance  $x$ , say  $N$ , is then Poisson distributed with mean, say  $v e^x = e^{x + \ln v}$ , where the parameter  $v$  expresses the sampling intensity. For the lognormal species abundance distribution,  $x$  is normally distributed among species with mean  $\mu$  and variance  $\sigma^2$ , whereas the log of the Poisson mean,  $x + \ln v$ , is normal with mean  $\mu + \ln v$  and the same variance  $\sigma^2$ . We define the Poisson lognormal distribution with parameters  $(\mu, \sigma^2)$  as the mixing distribution obtained when the log of the Poisson mean is normally distributed with mean  $\mu$  and variance  $\sigma^2$  (Grundy 1951, Bulmer 1974, O'Hara 2005, Golicher et al. 2006), imply-

ing that the mean is lognormally distributed. Accordingly, the number of individuals sampled from all  $s$  species in a community then constitutes a sample from the Poisson lognormal distribution with parameters  $(\mu + \ln v, \sigma^2)$ . For  $v=1$  this is the Poisson lognormal distribution with parameters  $(\mu, \sigma^2)$ , as follows

$$P(N = n; \mu, \sigma^2) = q(n; \mu, \sigma^2) = \int_{-\infty}^{\infty} h_n(\mu, \sigma^2, u) \phi(u) du,$$

where

$$h_n(\mu, \sigma^2, u) = \frac{\exp(u\sigma n + \mu n - e^{-(u\sigma + \mu)})}{n!},$$

and  $\phi(u)$  is the standard normal distribution.

Since  $s$  is unknown, we only consider the observed number of individuals for the species represented in the sample as proposed by Fisher et al. (1943). For a given sampling intensity  $v$ , the distribution of the number of individuals then follows the zero-truncated Poisson lognormal distribution

$$q(n; \mu + \ln v, \sigma^2) / [1 - q(0; \mu + \ln v, \sigma^2)],$$

defined for  $n = 1, 2, \dots$ . The maximum likelihood estimation of the parameters of this distribution was first derived by Bulmer (1974).

The similarity  $\rho$  between two communities (possibly different locations and different times) can be estimated from random samples from the communities, using the corresponding two-dimensional model. For two communities considered jointly the log abundances among species has the bivariate normal distribution with parameters, say  $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho)$ . The same sampling assumptions now lead to the bivariate distribution of the number of individuals of a species in the two samples,  $(N_1, N_2)$ , conditional on presence in at least one of them. This bivariate Poisson lognormal distribution takes the form

$$q(n_1, n_2; \mu_1 + \ln v_1, \sigma_1^2, \mu_2 + \ln v_2, \sigma_2^2, \rho) / [1 - q(0, 0; \mu_1 + \ln v_1, \sigma_1^2, \mu_2 + \ln v_2, \sigma_2^2, \rho)],$$

where the function  $q$  here is redefined for the two-dimensional case as

$$q(n_1, n_2; \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_{n_1}(\mu_1, \sigma_1^2, u) h_{n_2}(\mu_2, \sigma_2^2, v) f(u, v; \rho) du dv.$$

Here  $f(u, v; \rho)$  denotes the standard bivariate normal distribution with correlation  $\rho$ . Simulations from the bivariate lognormal distribution and corresponding Poisson lognormal distributions of counts are illustrated in Fig. 1.

From two samples with  $S \leq s$  different species observed in at least one of the samples, that is,  $S$  recordings of  $(N_1 = n_1, N_2 = n_2)$ , the log likelihood is the sum of  $\ln[q(n_1, n_2) / (1 - q(0, 0))]$  over the  $S$  observed species with inserted parameters. Hence, computation of the likelihood requires a large number of two-dimensional integrations, so the numerical maximization is a bit time-consuming.

Notice that the parameters  $\sigma_1^2, \sigma_2^2$  and  $\rho$  can be estimated without any knowledge of the unknown sampling intensities  $v_1$  and  $v_2$ . This is a very important observation because it makes it possible to perform a complete spatio-temporal analysis even when the sampling intensities vary and are

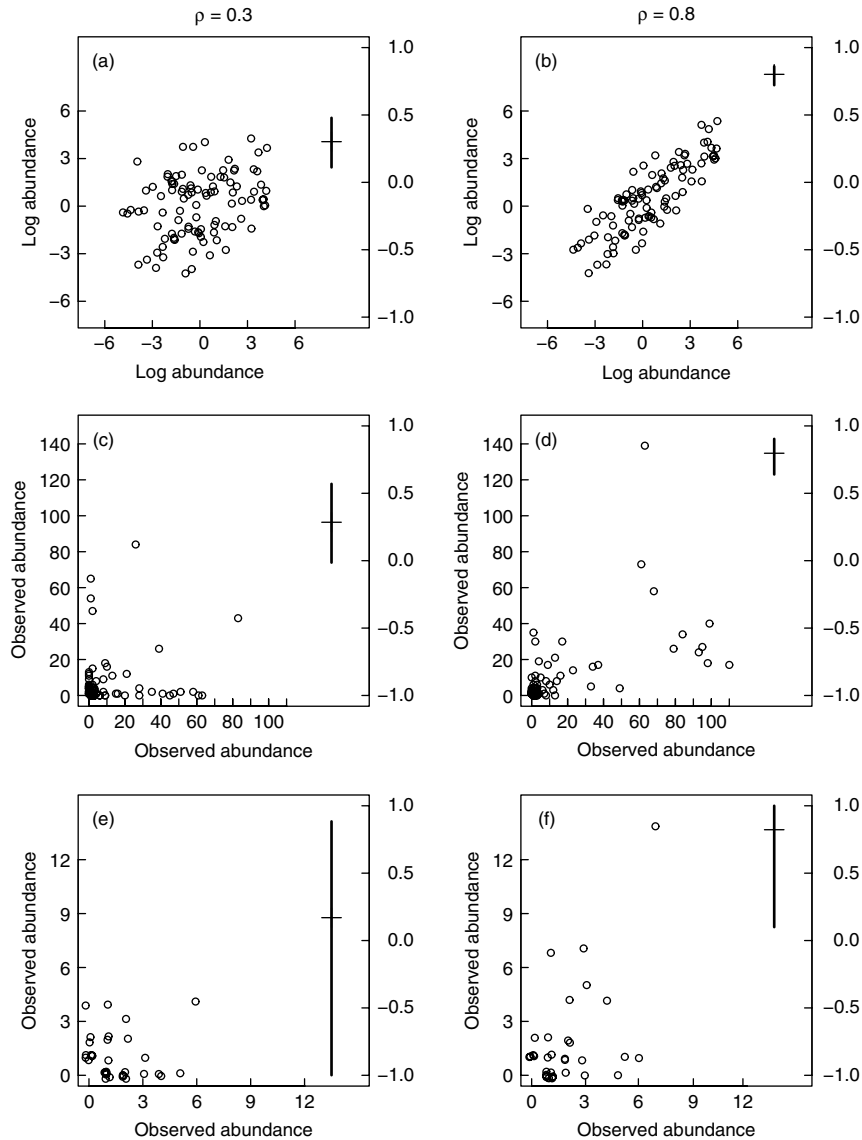


Figure 1. Simulated realizations from the bivariate lognormal distribution (a) and (b) and the bivariate Poisson lognormal distribution (c–f) for different correlations  $\rho$  between the log-abundances of the the species in the two communities. The sampling intensities  $v_1 = v_2$  in the bivariate Poisson log-normal distribution are adjusted so that 80% (middle row) and 30% (lower row) out of the 100 species in total are expected to be observed during the Poisson sampling process. The vertical line to the right in each panel show the interval covering 95% of estimated correlations  $\rho$  obtained by generating 500 realizations of the data and estimating  $\rho$  by maximum likelihood, and the horizontal line shows the median estimate. Other parameter values are  $\mu_1 = 0$ ,  $\mu_2 = 0$ ,  $\sigma_1^2 = 4$ ,  $\sigma_2^2 = 4$ .

unknown, which is often the case. The estimation can be performed using the R-package (R Development Core Team 2007) `poilog`. The mean values are confounded with the sampling intensities and cannot be estimated. However, the parameters  $\mu_i + \ln v_i$  of the Poisson lognormal distribution can be estimated, implying that estimates of the  $\mu_i$  can be found if the sampling intensities  $v_i$  are known.

Examples showing approximately unbiased estimation of  $\rho$  from samples of different sizes are illustrated in Fig. 2. A number of other simulations (see also Fig. 4) indicates that the correlation estimate is approximately unbiased for a realistic range of parameter values. Notice that the uncertainty seen in Fig. 2 only decreases very slowly as the number of individuals in the sample

increases. The reason for this is that the number of observations  $S$  entering the likelihood function is actually the number of observed species and not the number of individuals sampled. As the number of individuals sampled increases, the number of species sampled increases slowly as indicated in Fig. 2, and the standard deviation of the estimated correlation is roughly proportional to  $1/\sqrt{S}$ . Even for an infinite sample size we can not do better than to actually observe the underlying densities of individuals which is a sample from the bivariate lognormal distribution of size  $s$ , the number of species in the community. Hence, the variance of these estimators will not approach zero as the sample size approaches infinity as is usually the case in standard estimation theory.

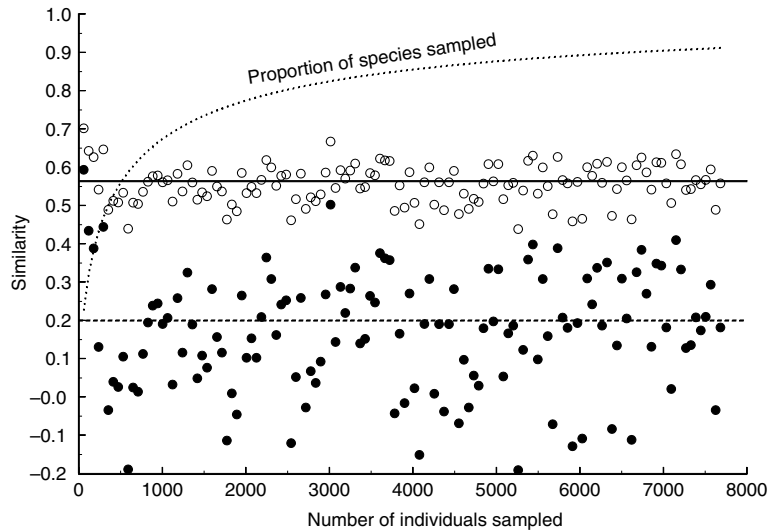


Figure 2. Simulations from the bivariate Poisson lognormal model with  $s=80$  and  $\sigma_1^2 = \sigma_2^2 = \mu_1 = \mu_2 = 0$  and  $\rho = 0.2$  for different sampling intensities giving different expected number of individuals and species in the sample. The filled circles are the maximum likelihood estimates of  $\rho$  fluctuating around the true value (dashed line). The open circles are the estimates of the Sørensen-index standardized by choosing  $p = q = 0.5$  fluctuating around the true value (solid line). These are estimated by first estimating  $\rho$  and then calculating the index using the expression for the Sørensen-index as a function of  $\rho$  given in the Appendix. The dotted line is the expected fraction of species represented in each sample.

## Deviations from the assumption of a lognormal abundance distribution

Although many communities seem to have species abundances following approximately the lognormal distribution, other models have been discussed widely in the literature. The deviation seen in practice is usually that there are more rare species than those predicted by log normality. This is the form one finds for the distribution of log abundance when the abundances follow a gamma distribution with shape parameter  $k$  smaller than one. The extreme case of Fisher's logarithmic series model with  $k=0$  corresponds to the degenerate distribution that is constant for large negative log abundances, representing an extreme, actually infinite, negative skewness. Also, the neutral model of Hubbell (2001) leads to distributions skewed to the left typically between the normal distribution and Fisher's model. To investigate how the correlation estimate performs when the distribution has this skewness we have performed simulations and estimation for a bivariate gamma model. This model can be obtained by defining two abundances ( $X, Y$ ) by  $X = W_1 + W_2$  and  $Y = W_2 + W_3$ , where the  $W_i$  are independent gamma distributed variables with scale parameter 1 (which is not interesting because it is confounded with sampling intensity) and shape parameters  $kr$  for  $W_2$  and  $k(1-r)$  for  $W_1$  and  $W_3$ . Then the marginal distributions of  $X$  and  $Y$  are gamma with shape parameter  $k$ , and  $\text{corr}(X, Y) = r$ . However, to compare to the lognormal model we must find the corresponding correlation between log abundances  $\text{corr}(\ln X, \ln Y)$ . This can be found by numerical integration, but is most easily computed by calculating the correlation from an extremely large simulated sample of  $(\ln X, \ln Y)$ . This gives approximately the log abundance correlation  $\rho$  as function of  $r$ . Then we have simulated count data from this model under random Poisson sampling. The counts then follow a bivariate

negative binomial distribution. From these simulated data we have estimated the correlation  $\rho$  using the bivariate Poisson lognormal. The results given in Fig. 3 show surprisingly small bias in the estimator, even for distributions that are heavily skewed to the left. This indicates that the estimation method based on the bivariate Poisson lognormal model is robust against the type of skewness one often finds in real communities.

## Relations to indices of similarity

The community correlation  $\rho$ , that can be estimated without any knowledge of the two sampling intensities involved, is also a measurement of similarity which has been used to perform spatial as well as temporal analysis of community dynamics (Engen et al. 2002, 2008, Lande et al. 2003). We now proceed to analyze the relation between  $\rho$  and the Jaccard-index calculated directly from the samples which is

$$J = \frac{A}{A + B + C},$$

where  $A$  is the number of species present in both samples, while  $B$  and  $C$  are numbers of species present only in one or the other. The Sørensen-index is defined as

$$L = \frac{2A}{2A + B + C}.$$

The relations between these indices are simply  $L = 2J / (J + 1)$  and  $J = L / (2 - L)$ . Thus, they are actually equivalent in the sense that the value of each one of them uniquely determines the value of the other.

Samples of species from communities collected at the same or different locations, possibly at different times, are

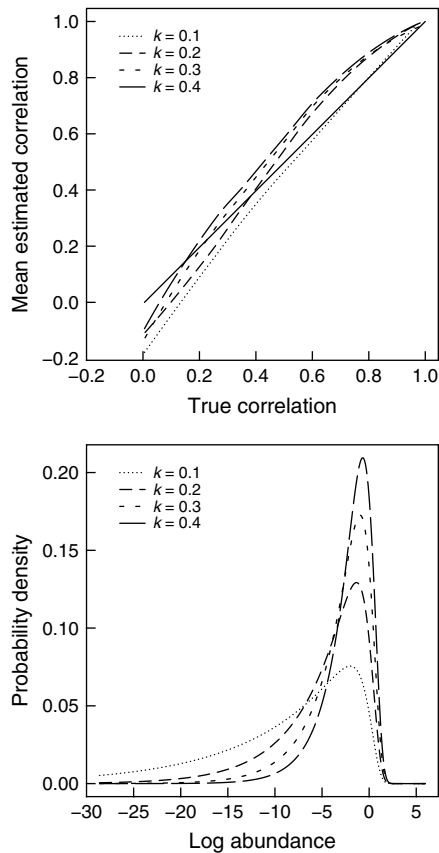


Figure 3. The upper panel shows estimated correlation between log abundances, found by fitting the bivariate Poisson lognormal model to abundance data simulated from different bivariate gamma models for abundances (described in the main text), against the true correlation between log abundances. In the bivariate gamma distribution with shape parameter  $k$  and scale parameter 1 the expected fraction of species  $q$  observed in the sample is given by  $q = 1 - (1 + \nu)^{-k}$  and the sampling intensity  $\nu$  that gives a certain fraction of observed species is  $\nu = (1 - q)^{-1/k} - 1$ . For each values of  $k$  sampling intensities were chosen so that a fraction  $q = 0.6$  of the species in the community was expected to be observed in samples. For correlations  $r = 0.01, 0.02, \dots, 1$  in the bivariate gamma distribution we estimated the corresponding correlation between log abundance from 2 000 000 samples. Estimates of correlations in log abundance by fitting the bivariate Poisson lognormal distribution were based on 500 samples. Dashed lines show estimated correlations as functions of true correlations for different values of  $k$ , smoothed by polynomial regression. The solid line indicates equality between the true correlation and estimated correlation. The lower panel shows the marginal distributions of log abundance for the 4 different gamma distributions used in the upper panel which have shape parameters 0.1 (largest negative skewness and largest deviation from the normal distribution), 0.2, 0.3 and 0.4 (smallest negative skewness, closest to the normal distribution).

always different (Colwell and Coddington 1994, Lande 1996, Gotelli and Colwell 2001). Even parallel samples from the same locality taken at the same time differ because many species tend to be rare in the samples so that their presence is a stochastic event (Magurran 2004, Mao and Colwell 2005). Thus, the recording of a single individual or species will depend on the sampling effort as illustrated in Fig. 1. In addition, abundant species often show large

differences in abundance in space due to spatial heterogeneity, as well as fluctuations through time generated by environmental and demographic stochasticity (Engen et al. 2002, Lande et al. 2003, Engen 2007a, b, Magurran 2007, Engen et al. 2008). Thus, samples only contain a fraction of the species actually present in the community. When the Jaccard- and Sørensen-indices defined for the total communities are estimated by replacing  $A$ ,  $B$  and  $C$  by the corresponding number of species recorded in the samples, their values (or the mean value under sampling) will vary considerably with sample size and sampling effort (Fig. 1, 2), making comparisons among characteristics of communities difficult.

In the bivariate lognormal species abundance model the bivariate log abundance of a species in the two communities,  $(X, Y)$ , is assumed to be generated from a bivariate normal distribution. The total community of  $s$  species is accordingly assumed to be a sample of size  $s$  from the bivariate normal density. The corresponding standardized log abundances  $U = (X - \mu_1)/\sigma_1$  and  $V = (Y - \mu_2)/\sigma_2$  then follow the standard bivariate normal distribution with correlation  $\rho$ , that is, the marginal distributions have zero means and unit variances.

Under the assumption of Poisson sampling the probability that a species with log abundance  $x$  is present in a sample under sampling intensity  $\nu$  is  $1 - \exp(-\nu e^x)$  and the expected fraction of species present in a sample is the expected value of this probability with respect to the distribution of  $x$  among species in the community. However, since the correlation  $\rho$  is defined as a property of the community unaffected by the stochasticity in the sampling, we should rather define “presence” without reference to a sampling distribution in order to compare the measurements. For a given sampling procedure, the presence of a species depends on the sampling intensity. If we for example define “presence” as “the probability that the species is present in the sample is larger than 1/2”, this is equivalent to defining “presence” as “the abundance exceeds a given threshold”. In the above example of Poisson sampling, using probability 1/2 in the definition, this threshold is  $\ln(\ln 2) - \ln(\nu)$ . For a different sampling distribution, such as the negative binomial or Poisson lognormal, this threshold would be different, but still a decreasing function of  $\nu$ . For any sampling model in which the probability of observing a species increases with its abundance the “presence” defined as above (with 1/2 replace by any probability) is equivalent to “ $x$  exceeding some threshold”. Therefore, in order to compare the correlation measure to the Jaccard- and Sørensen-indices with no reference to particular sampling distributions we define “presence” as “abundance exceeding some threshold”. Decreasing the threshold then increases the number of species “present” corresponding to what happens in general when sampling intensities increase and more rare species occur in the sample.

For the bivariate lognormal species abundance model it is mathematically most convenient to define thresholds referring to the standardized log abundances. Hence, for a species with log abundance  $X$  we consider the event  $U = (X - \mu_x)/\sigma_x > z_u$ . The expected fraction of species exceeding the threshold  $z_u$  for  $U$ , that is, the fraction of species defined as present, is then  $P(U > z_u) = 1 - \Phi(z_u)$ , where  $\Phi$  is the standard cumulative normal distribution.

Hence, the threshold  $z_u$  is comparable to a sampling effort that reveals a fraction  $p = 1 - \Phi(z_u)$  of the species abundance distribution on the log scale (as well as on the absolute scale). Hence, specifying this threshold is equivalent to specifying some sampling intensity, actually the intensity corresponding to the given fraction of species being “present”. Correspondingly, for the other community, we define a threshold  $z_v$  for  $V = (Y - \mu_y) / \sigma_y$ . The expected fraction of species with standardized abundance above  $v$  in the second sample is then  $q = 1 - \Phi(z_v)$ .

For the two-dimensional model the distribution of  $(U, V)$  among species is now the standard bivariate normal distribution  $f(u, v; \rho)$  with correlation  $\rho$ . Writing  $s$  for the number of species, the expectation of quantities entering the expression for the Jaccard- and Sørensen-indices are  $A = s \int_{z_u}^{\infty} \int_{z_v}^{\infty} f(u, v; \rho) dudv$ ,  $B = s \int_{z_u}^{\infty} \int_{-\infty}^{z_v} f(u, v; \rho) dudv$  and  $C = s \int_{-\infty}^{z_u} \int_{z_v}^{\infty} f(u, v; \rho) dudv$ . When these integrals are plugged into the expressions for  $J$  and  $L$ , the parameter  $s$  disappears, so that the indices, say  $J_{p,q}$  and  $L_{p,q}$ , are simply functions of the community correlation  $\rho$  only. Now, writing  $G(z_u, z_v; \rho) = P(U > z_u, V > z_v)$  and using the fact that the distribution of  $U$  given  $V = v$  is normal with mean  $\rho v$  and variance  $(1 - \rho^2)$ , the two-dimensional integral defining  $G(z_u, z_v; \rho)$  can be written as the univariate integral

$$G(z_u, z_v; \rho) = \int_{z_v}^{\infty} \left[ 1 - \Phi \left( \frac{z_u - \rho v}{\sqrt{1 - \rho^2}} \right) \right] \phi(v) dv,$$

where  $\phi(v)$  is the standard normal density. Writing  $P(U > z_u, V < z_v) = P(U > z_u, -V < -z_v)$  and using the fact that  $(U, -V)$  is standard bivariate normal with correlation  $-\rho$ , we find  $P(U > z_u, V < z_v) = G(z_u, -z_v; -\rho)$ , and in the same way  $P(U < z_u, V > z_v) = G(-z_u, z_v; -\rho)$ . Hence, the indices referring to expected fractions  $p$  and  $q$  of species being revealed by the samples, can be written as

$$J_{p,q} = \frac{G(z_u, z_v; \rho)}{G(z_u, z_v; \rho) + G(-z_u, z_v; -\rho) + G(z_u, -z_v; -\rho)}$$

and

$$L_{p,q} = \frac{2G(z_u, z_v; \rho)}{2G(z_u, z_v; \rho) + G(-z_u, z_v; -\rho) + G(z_u, -z_v; -\rho)}.$$

Assuming  $p = q$ , we see from Fig. 4 that the Sørensen-index  $L$  for two communities containing an equal number of species is dependent on the sampling intensity as well as the correlation of the log abundances  $\rho$ .

The above integrals  $G(z_u, z_v; \rho)$  must be computed numerically except in the special case when one half of the species in the communities are expected to be sampled, that is,  $p = q = 0.5$ . Analytical solutions for  $J_{0.5,0.5}$  and  $L_{0.5,0.5}$  are given in the Appendix.

Simulation of samples based on two communities with an equal number of species, but different correlation of the log-abundances, shows that the mean estimate of the correlation-index is, in contrast to the Sørensen-index, independent of the proportion of species sampled (Fig. 5), although the uncertainties for low sampling intensities may

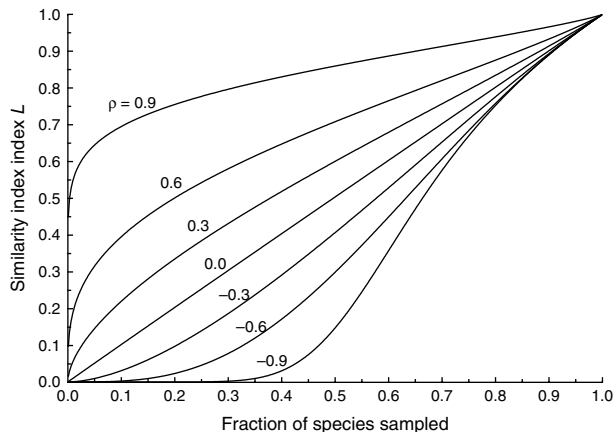


Figure 4. Sørensen's similarity index  $L$  as a function of mean fractions of species  $p = q$  exceeding the thresholds for different values of the correlations between the log-abundances of species in the two communities  $\rho$ .

be large. The estimate of the abundance-based analogy of Sørensen's-index  $L$  defined by Chao et al. (2005) also tends to decrease with decreasing sampling intensity.

### Example: temperate bird communities

In this section we will illustrate the practical application of our method by making two comparisons of bird communities. In a bird community in East Forest, Surrey, outside London in southern England (Gaston and Blackburn 2000) the correlation of the log-abundances at two different years was high (Fig. 6a,  $\hat{\rho} = 1.00$  with 95% confidence interval  $[1.00, 1.00]$ ), showing large similarity over time in community composition. Notice however, that the counts are not on a straight line even if the correlation estimate is 1. This is because it is the correlation between the underlying species abundances that is one, not the correlation between the counts which is somewhat smaller due to sampling effects. In contrast, the similarity between two bird communities located ca 350 km apart in central Norway (Hogstad 1967, 1968) was substantially smaller (Fig. 6b,  $\hat{\rho} = 0.60$  with 95% confidence interval  $[-0.28, 0.99]$ ). Notice that the value of the correlation is hard to judge just from inspection of the scatter plot of observed counts due to the effect of Poisson sampling.

### Discussion

It is well known that estimates of indices of similarity based on presence-absence data are very sensitive to sampling effort (Lande 1996, Gotelli and Colwell 2001). As sample sizes increase, an increasing number of species will be present in both samples and the indices tend to approach their theoretical upper value. Indices based on species abundances rather than presence and absence are generally easier to estimate in a way that corrects for sampling effort. For example, this is the case for the classical Morisita-index of similarity (Morisita 1959), which is closely related to Simpson's diversity index (Simpson 1949). This index is a

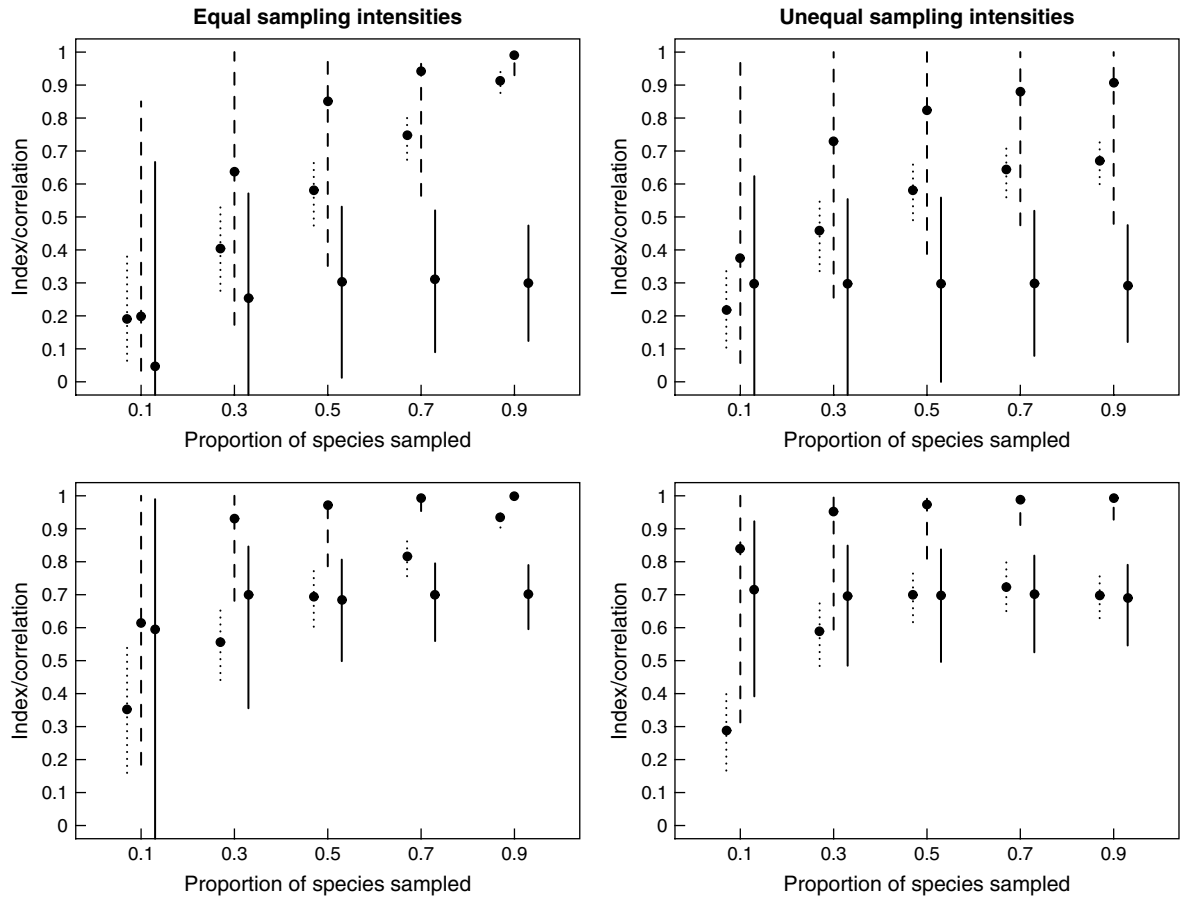


Figure 5. The 95% interval estimates of the incidence-based Sørensen-index based on simple species counts (dotted lines), the abundance based Sørensen-index (dashed lines) and the correlation  $\rho$  (solid lines) estimated from the bivariate Poisson lognormal distribution. Data was simulated from the bivariate Poisson lognormal distribution with parameters  $\mu_1 = 0$ ,  $\mu_2 = 0$ ,  $\sigma_1^2 = 7$ ,  $\sigma_2^2 = 7$ , and  $\rho$  was 0.3 and 0.7 in the upper row and lower row respectively. The total number of species in the communities was 150, and the sampling intensities  $v_1 = v_2$  in the left column were adjusted so that the expected fraction of species observed was 0.1, 0.3, 0.5, 0.7 and 0.9 respectively. In the right column we varied the sampling intensity  $v_1$  such that the expected fraction of species observed in sample 1 was 0.1, 0.3, 0.5, 0.7 and 0.9 respectively, whereas sampling intensity  $v_2$  was held constant such that the expected fraction of species observed in sample 2 was 0.5, thus giving unequal expected number of species observed in the samples.

dominance index, and is therefore essentially only affected by the few most common species in the communities.

Here we use the bivariate lognormal species abundance model, assuming that the two communities have the same species in agreement with several stochastic spatio-temporal community models (Engen and Lande 1996a, Engen et al. 2002, Lande et al. 2003, Engen 2007a, b). In these models there is no lower bound for the log abundances, so in practice species may be unobservable in one site while being observable in the other. Using the correlation as a measurement of similarity, we obtain an abundance-based index of similarity that, contrary to the Morisita (1959)-index, is equally affected by common and rare species. This model enables us to express the most well known similarity indices based on presence-absence as functions of the correlations between the communities when presence of species is interpreted as abundance exceeding a given threshold. Fortunately, this correlation can be estimated by maximum likelihood even if the sampling intensities at the two locations are different and unknown (Fig. 5). Hence, the value of presence-absence indices referring to any sample sizes can also be estimated. Alternatively, if the

indices are defined using expected number of species in the samples relative to a given sample distribution, such as the Poisson, the index is basically still a function of the correlation (Fig. 5, 7), but it may also depend slightly on other parameters. However, by fitting the bivariate Poisson lognormal model to the data, all parameters can be estimated and values of the indices relative to any given sample sizes can be evaluated.

Many proposed indices of similarity based on the counts  $A$ ,  $B$  and  $C$  are equivalent. Koleff et al. (2003) and Chao et al. (2006) list a number of indices of this type, many of which can be expressed as functions of the correlation in the bivariate lognormal model. If we physically consider  $A$ ,  $B$  and  $C$  to have “dimension”  $d$ , indices can be expressed by  $\rho$  if they have dimension 1. The Jaccard- and Sørensen-indices are fractions with dimension  $d$  in the numerator and denominator, thus having dimension 1. A number of indices are of a similar type, with the same dimension in the numerator and denominator, giving dimension 1. In all these indices, the species number disappears as for the Jaccard- and Sørensen-indices so that the index is a function of the probabilities  $G(z_u, z_v; \rho)$  only.

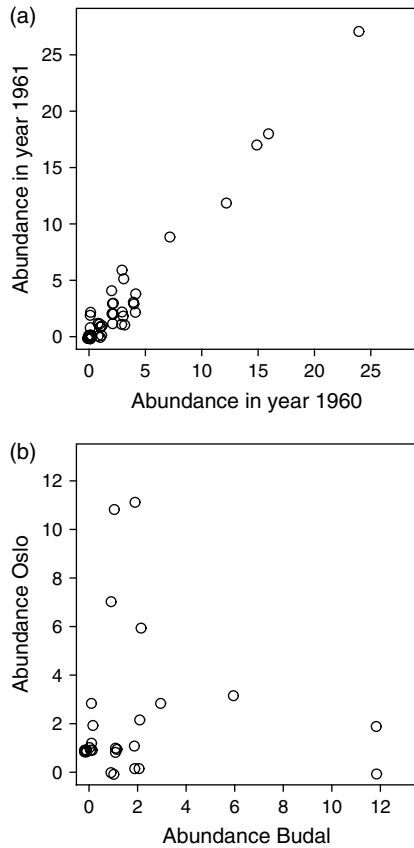


Figure 6. The correlation between log-abundances of bird species in Eastern Wood, Surrey (Gaston and Blackburn 2000) in the years 1960 and 1961 (a), and the correlation between a bird community outside Oslo in southern Norway in the year 1966 (Hogstad 1967) and in Budal in central Norway (Hogstad 1968) in the year 1968 (b). The parameter estimates were  $\mu_1 + \ln v_1 = 0.57 [-0.42, 1.11]$ ,  $\mu_2 + \ln v_2 = 0.55 [-0.37, 1.14]$ ,  $\sigma_1^2 = 1.33 [0.39, 2.96]$ ,  $\sigma_2^2 = 1.50 [0.49, 3.05]$  and  $\rho = 1 [1, 1]$  in (a), and  $\mu_1 + \ln v_1 = -1.16 [-3.77, 0.04]$ ,  $\mu_2 + \ln v_2 = -0.56 [-2.89, 0.56]$ ,  $\sigma_1^2 = 3.03 [0.66, 8.26]$ ,  $\sigma_2^2 = 1.94 [0.27, 5.33]$  and  $\rho = 0.60 [-0.28, 0.99]$  in (b). Subscript 1 corresponds to year 1960 in (a) and Budal in (b). Numbers inside square brackets denote the lower and upper limits for 95% confidence intervals obtained by parametric bootstrapping. We calculated a goodness of fit statistic  $\hat{c}$  based on the distribution of the deviance calculated from the fitted model divided by the deviance calculated from 1000 bootstrap estimates (see Connolly et al. 2009 for details). This test statistic normalizes model deviance relative to the expected level of deviance due to random sampling and a value of  $\hat{c} = 1$  indicates that that estimated model's lack of fit is equal to the lack of fit expected due to random sampling from the model. Estimates (mean and 95% confidence limits) were  $\hat{c} = 0.957 [0.67, 1.53]$  and  $\hat{c} = 1.13 [0.72, 2.01]$ , indicating that there is no strong evidence for lack of fit of the Poisson lognormal model.

Because the fractions  $p$  and  $q$  are approximately the fraction of species found in the samples, they can be estimated as one minus the zero-term of the univariate Poisson lognormal distribution with estimated values of  $\mu_i + \ln v_i$  and  $\sigma_i^2$  plugged in (Bulmer 1974). This enables estimation of the Jaccard- and Sørensen-indices for the actual sample size using the relation derived earlier. However, more interestingly, we can now estimate the indices defined by any choice of  $p$  and  $q$  using the general

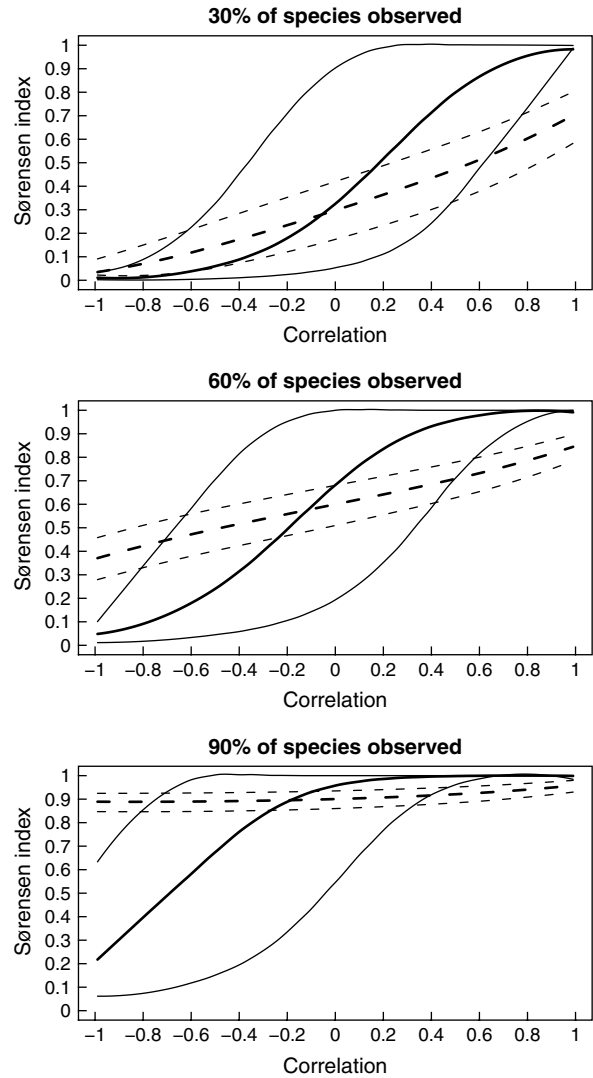


Figure 7. The incidence based Sørensen-index (dashed lines) and the abundance based (Chao et al. 2006) Sørensen-index (solid lines) estimated from samples from the bivariate Poisson lognormal distribution with parameters  $\mu_1 = 0$ ,  $\mu_2 = 0$ ,  $\sigma_1^2 = 7$ ,  $\sigma_2^2 = 7$ , and  $\rho$  varying from  $-1$  to  $1$  (x-axis). The sampling intensities  $v_1 = v_2$  were chosen to reflect varying fractions of species observed in the samples. The thick lines show medians whereas the thin lines show the 2.5 and 97.5% quantiles of the distribution of parameter estimates obtained by a large number of repeated samples.

expression for these indices as functions of  $\rho$ . On the other hand, since these are just increasing functions of  $\rho$ , nothing is really gained by transforming from  $\rho$  to one of these indices.

Our approach differs from the classical indices just based on the presence and absence by using a parametric model that utilizes the abundances of the species and accounts for sampling. Chao et al. (2005, 2006) also included the relative species abundances in a derivation of an index of community similarity that had a similar mathematical form as the Jaccard- and Sørensen-indices. For two communities let  $U_1$  and  $U_2$  be the relative abundances of the shared species. Chao et al.'s analogy to the Jaccard-index is then

$$J_{chao} = \frac{U_1 U_2}{U_1(1 - U_2) + U_2(1 - U_1) + U_1 U_2}$$

$$= \frac{U_1 U_2}{U_1 + U_2 - U_1 U_2}.$$

Although this is a mathematical analogy to the Jaccard-index, it is a very different measure of similarity. A striking similarity is that it equals one if all species are shared, but a large difference is that Chao's index depends on the abundances  $U_i$  of shared species, treating all shared species as a group. When these relative abundances are given, the value of the index is not affected by the species numbers, neither the shared ones nor those present in only one community (Fig. 7). In addition, it is not influenced by the distribution of the abundances of the species that are not present in both samples. If one individual is chosen at random from each of two communities  $i$  (labeled 1, 2), let  $C_i$  denote the event that they belong to the shared species in community  $i$ . The Chao-index can then alternatively be written as  $J_{chao} = P(C_1 \cap C_2) / P(C_1 \cup C_2) = P(C_1 \cap C_2 | C_1 \cup C_2)$ . Neither of these three expressions for the index make it easy to interpret in practice. If two communities have index 0.4, while two others have 0.6, these values alone give little information about the true similarities because it is not based on single species abundances, but only the total abundance of all shared species. Hence, it may not make the appropriate distinction between communities that are very different with respect to abundances (Fig. 5, 7). If two communities are equivalent (same species and same relative abundances), then  $J_{chao} = 1$ . This is because the index is constructed for a different reason than our correlation  $\rho$ , focusing on proportion of individuals belonging to species that are shared across communities. Therefore,  $J_{chao}$  still takes the same value if the species with large abundance in one community are those with small abundance in the other (Fig. 7), corresponding to a negative correlation. The index may therefore be inappropriate for comparison of communities that basically have the same set of species, which is the rule rather than the exception in analysis of  $\beta$ -diversity in communities of species distributed continuously in space or in analysis of temporal fluctuations in community composition at a given location.

One of the most important applications of indices of similarity is in analyses of spatial and temporal variation in species diversity of communities. In particular, when measuring temporal similarity, that is, the similarity between the community seen at a given location at two different times, the same species are likely to have a non-zero abundance during both sampling processes. This is because extinctions are generally rare. However, with many rare species, samples may still appear as rather different. Actually, due to the stochasticity in the sampling process, samples may look rather different even if the communities are identical. This is illustrated in Fig. 6a, where samples are taken in 1960 and 1961. The correlation estimate appears to be one, even if the points are not on a straight line. The deviation from a line is not large enough to conclude that it is not just a sampling effect in two perfectly correlated communities. In Fig. 6b, however, there is a large spatial distance between the samples, as well as a

temporal distance of 2 yr, giving an estimate of 0.60. Again, similarity may seem smaller by inspection of the graph, illustrating the difficulty in estimation, which is here overcome by using the maximum likelihood estimate assuming Poisson sampling. In this example there is a large group of species that are common in both samples as well as a group of species that are rare in both. Accounting for sampling effects, this leads to the relatively large value of  $\rho$ . More detailed analysis of temporal as well as spatial analysis of communities based on the correlation measure are given in Engen et al. (2002, 2008), Lande et al. (2003), and Walla et al. (2004). These applications utilized properties of the underlying spatial and temporal processes that actually generate multivariate normal distributions of log abundance in space and time. This enables partitioning of the variance parameter in the lognormal species abundance distribution into components of species diversity due to permanent differences between species, heterogeneity in the landscape and sampling effects, as well as components due to environmental stochasticity in the temporal process of each single species. The statistical analysis based on estimating correlations between communities makes it possible to estimate all these components. Such a thorough analysis of  $\beta$ -diversity can hardly be done with other indices because they lack the link to multivariate normal theory that makes this decomposition possible.

One interesting yardstick for presence-absence indices is based on the sample sizes where half the species in the communities are expected to be present in the sample, i.e.  $z_u = z_v = 0$ . For the lognormal model this is the sample size with the largest number of species appearing as rare in the samples, hence it is the sample size likely to be most sensitive to stochastic presence and absence of species. However, using the threshold definition, all definitions referring to specific fractions of species counted, as well as all the indices listed by Chao et al. (2005) are equivalent as long as the description of the two communities can be given approximately by the bivariate lognormal species abundance model.

Using the bivariate lognormal species abundance distribution is, in principle, not different from using the normal and bivariate normal distribution as approximations in statistics in general. The product moment correlation (Pearson's correlation) is in any case a measurement of covariation between variables, so the estimate obtained from the bivariate Poisson lognormal is likely to be a relevant description of similarity even if the bivariate lognormal model only is an approximation to the real species abundance distribution. We have checked the robustness of the method by applying it when the abundances of two communities follow a bivariate gamma distribution. The method performs surprisingly well even for marginal distributions of log abundance that are far from normal with negative skewness which is sometimes realistic.

*Acknowledgements* – This project has been supported by the Norwegian Research Council, project 159571/V40, Stochastic dynamics of bird communities. We want to thank Russell Lande for interesting discussions and valuable comments to previous versions of the manuscript.

## References

- 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.
- Chao, A. et al. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. – *Biometrics* 62: 361–371.
- Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity by extrapolation. – *Phil. Trans. R. Soc. B* 345: 101–118.
- Connolly, S. R. et al. 2009. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. – *Ecology* 90: 3138–3149.
- Engen, S. 1974. On species frequency models. – *Biometrika* 61: 263–270.
- Engen, S. 1978. Stochastic abundance models with emphasis on biological communities and species diversity. – Chapman and Hall.
- Engen, S. 2001. A dynamic and spatial model with migration generating the log-Gaussian field of population densities. – *Math. Biosci.* 173: 85–102.
- Engen, S. 2007a. Heterogeneity in dynamic species abundance model: the selective effect of extinction processes. – *Math. Biosci.* 210: 490–507.
- Engen, S. 2007b. Heterogeneous communities with lognormal species abundance distribution: species–area curves and sustainability. – *J. Theor. Biol.* 249: 791–803.
- Engen, S. and Lande, R. 1996a. Population dynamic models generating the lognormal species abundance distribution. – *Math. Biosci.* 132: 169–184.
- Engen, S. and Lande, R. 1996b. Population dynamic models generating species abundance distributions of the gamma type. – *J. Theor. Biol.* 178: 325–331.
- Engen, S. et al. 2002. Analyzing spatial structure of communities by the two-dimensional Poisson lognormal species abundance model. – *Am. Nat.* 160: 60–73.
- Engen, S. et al. 2008. Assessment of species diversity from species abundance distributions at different localities. – *Oikos* 117: 738–748.
- 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: 42–58.
- Gaston, K. J. and Blackburn, T. M. 2000. Pattern and processes in macroecology. – Blackwell.
- 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.
- 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.
- Hogstad, O. 1967. Seasonal fluctuations in bird populations within a forest area near Oslo (southern Norway) in 1966–67. – *Norw. J. Zool.* 15: 81–96.
- Hogstad, O. 1968. Breeding bird populations in two subalpine habitats in the middle of Norway during the years 1966–68. – *Norw. J. Zool.* 17: 81–91.
- Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.
- Jaccard, P. 1912. The distribution of the ora in the alpine zone. – *New Phytol.* 11: 37–50.
- Koleff, P. et al. 2003. Measuring beta diversity for presence–absence data. – *J. Anim. Ecol.* 72: 367–382.
- 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.
- Leigh, E. G. 2007. Neutral theory: a historical perspective. – *J. Evol. Biol.* 20: 2075–2091.
- Loreau, M. and de Mazancourt, C. 2008. Species synchrony and its drivers: neutral and nonneutral community dynamics in fluctuating environments. – *Am. Nat.* 172: E48–E66.
- Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.
- Magurran, A. E. 2007. Species abundance distributions over time. – *Ecol. Lett.* 10: 347–354.
- 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.
- May, R. M. 1975. Patterns of species abundance and diversity. – In: Cody, M. L. and Diamond, J. M. (eds), *Ecology and evolution of communities*. Harvard Univ. Press, pp. 81–120.
- 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.
- Morisita, M. 1959. Measuring the dispersion of individuals and analysis of the distributional patterns. – *Mem. Fac. Kyushu Univ. Ser. E (Biol.)* 2: 215–235.
- Mutshinda, C. M. et al. 2008. Species abundance dynamics under neutral assumptions: a Bayesian approach to the controversy. – *Funct. Ecol.* 22: 340–347.
- Mutshinda, C. M. et al. 2009. What drives community dynamics? – *Proc. R. Soc. B* 276: 2923–2929.
- 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.
- Peet, R. K. 1974. The measurement of species diversity. – *Annu. Rev. Ecol. Syst.* 5: 285–307.
- Pielou, E. C. 1975. *Ecological diversity*. – Wiley.
- Preston, F. W. 1948. The commonness and rarity of species. – *Ecology* 29: 254–283.
- Preston, F. W. 1960. Time and space and the variation of species. – *Ecology* 41: 611–627.
- Preston, F. W. 1962. The canonical distribution of commonness and rarity. – *Ecology* 43: 185–215, 410–432.
- Preston, F. W. 1980. Non-canonical distributions of commonness and rarity. – *Ecology* 61: 88–97.
- R Development Core Team 2007. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria, <[www.R-project.org](http://www.R-project.org)>.
- Sheppard, W. F. 1898. On the application of the theory of error to cases of normal distributions and normal correlation. – *Phil. Trans. A* 192: 101–167.
- Simpson, E. H. 1949. Measurement of diversity. – *Nature* 163: 688.
- Sørensen, T. A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. – *Kgl Danske Vidensk. Selsk. Biol. Skr.* 5: 1–34.
- Tokeshi, M. 1993. Species abundance patterns and community structure. – *Adv. Ecol. Res.* 24: 111–186.
- Ugland, K. I. and Gray, J. S. 1982. Lognormal distributions and the concept of community equilibrium. – *Oikos* 39: 171–178.
- Walla, T. R. et al. 2004. Modeling vertical beta-diversity in tropical butterfly communities. – *Oikos* 107: 610–618.
- Whittaker, R. H. 1970. *Communities and ecosystems*. – Macmillan.
- Whittaker, R. H. 1972. Evolution and measurement of species diversity. – *Taxon* 21: 213–251.

## Appendix

### Mean log abundance as threshold

The integrals defining  $G(z_u, z_v; \rho)$  have to be computed numerically except in the case when the log abundance thresholds are the mean values of the marginal distribution, which is equivalent to saying that half of the species are expected to be counted. The indices then depend on  $G(0, 0; \rho)$ , which is the probability that the standardized log abundance of a species is positive in both communities. This is a classical problem in the analysis of the bivariate normal distribution first solved by Sheppard (1898), showing that the probability that the standardized bivariate normal variate takes a value in the first quadrant is  $1/4 + \arcsin(\rho)/(2\pi)$ . Hence  $G(0, 0; \rho) = 1/4 + \arcsin(\rho)/(2\pi)$ , and  $G(0, 0; -\rho) = 1/4 - \arcsin(\rho)$  giving

$$J^*(\rho) = \frac{1 + 2 \arcsin(\rho)/\pi}{3 - 2 \arcsin(\rho)/\pi},$$

where  $J^* = J_{0.5,0.5}$ . Since  $\arcsin(-1) = -\pi/2$ ,  $\arcsin(0) = 0$  and  $\arcsin(1) = \pi/2$  we see that  $J^*(-1) = 0$ ,  $J^*(0) = 1/3$  and  $J^*(1) = 1$ .

The corresponding Sørensen-index takes the simpler form

$$L^*(\rho) = 1/2 + \arcsin(\rho)/\pi$$

where  $L^* = L_{0.5,0.5}$ , with special values  $L^*(-1) = 0$ ,  $L^*(0) = 0.5$  and  $L^*(1) = 1$ . Expressing  $\rho$  by the indices we find

$$\rho = \sin \left[ \frac{\pi}{2} (2L^* - 1) \right] = \sin \left[ \frac{\pi}{2} \left( \frac{3J^* - 1}{J^* + 1} \right) \right].$$