Ecological Informatics 4 (2009) 76 –82

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

On the distribution of plant abundance data

Christian Damgaard *

Department of Terrestrial Ecology, NERI, University of Aarhus, Vejlsøvej 25, 8600 Silkeborg, Denmark

a r t i c l e i n f o

Article history:
Received 10 January 2008
Received in revised form 6 February 2009
Accepted 7 February 2009

Keywords:
Hierarchical sampling
Pin-point
Point-intercept
Plant cover
Spatial scale
Zero-in ated and bounded beta-distribution
Zero-in ated generalised binomial distribution

a b s t r a c t

Plant abundance data are often analysed using standard statistical procedures without considering their distributional features and the underlying ecological processes. However, plant abundance data, e.g. when measured in biodiversity monitoring programs, are often sampled using a hierarchical sampling procedure, and since plant abundance data in a hierarchical sampling procedure are typically both zero-in ated and over-dispersed, the use of a standard statistical procedure is sub-optimal and not the best possible practice in the modelling of plant abundance data. Two distributions (the zero-in ated generalised binomial distribution and the zero-in ated bounded beta distribution) are suggested as possible distributions for analysing either discrete, continuous, or ordinal hierarchically sampled plant cover data.

1. Introduction

Plant abundance is often described by the cover, i.e. the relative area covered by different plant species in a small plot. Plant cover is not biased by the size and distributions of individuals, and is an important and often measured characteristic of the composition of plant communities (Kent and Coker, 1992). Plant cover data may be used to classify the studied plant community into a vegetation type, to test different ecological hypothesis on plant abundance, and in gradient studies, where the effects of different environmental gradients on the abundance of speci c plant species are investigated (Austin, 2007). Such gradient studies may be used in the prognostic modelling of the effects of global warming, nitrogen deposition on plant community dynamics and, consequently, to predict the fate of speci c ecosystems in a changing environment (Guisan and Thuiller, 2005; Ibanez et al., 2006).

Plant abundance data are often analysed using standard statistical procedures without considering their distributional features. However, it is the claim of this study that the use of standard statistical procedures is sub-optimal and not the best possible practice in the modelling of plant cover data. In the process of developing a best possible practice of analysing plant cover data, it is important to specify stochastic models that describe the distribution of plant cover data from different sampling schemes, i.e. models that capture the most important ecological processes at different spatial scales. In recent years, there have been a number of articles dealing with the distribution of ecological data and, particularly, the problem of excess zeroes with ecological count data (e.g. Hall, 2000; Rathbun and Fei, 2006; Ver Hoef and Jansen, 2007; Chen et al., 2008). This trend is encouraging and signals the increasing awareness in the ecological community of the importance of using "tailored" distributions for the ecological data, rather than forcing the ecological data into standard distributions.

Plant cover data, e.g. when measured in biodiversity monitoring programs, are often sampled using a hierarchical sampling procedure, where several plots are sampled from a number of different sites. A relevant stochastic model of plant cover data from such hierarchical sampling will have to consider two important characteristics of the distribution of plant species: i) plant species do not occur everywhere possible and the data will, consequently, be zero-in ated, i.e. in some sites a speci c plant species may be totally absent due to random extinction events and/or limited possibility of the plant to colonise the habitat (MacArthur and Wilson, 1967; Rees et al., 2001; Leibold et al., 2004; Cordonnier et al., 2006), ii). If a plant species is present at the sites, the abundance of different plant species generally displays aggregated spatial patterns within the site due to e.g. the size of the plant, clonal growth, and limited seed dispersal, and plant cover data will typically be over-dispersed relative to random expectations (Pacala and Levin, 1997; Herben et al., 2000; Stoll and Weiner, 2000).

Both the large scale spatial process of extinction and recolonisation that leads to zero-in ated plant abundance data and the small-scale spatial process of plant growth and limited dispersal that leads to aggregated spatial patterns of plant abundance, result in an augmented variance compared to random expectations. Due to this augmented variance, there has been an unfortunate history in plant ecology i) re-coding plant abundance data into presence – absence data, thereby throwing away ecological information, ii) not reporting measured plant

* Tel.: +45 89201400; fax: +45 89201414.
E-mail address: cfd@dmu.dk.

1574-9541/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi: 10.1016/j.ecoinf.2009.02.002

abundance data in biodiversity monitoring programs, and iii) avoiding measuring plant abundance when setting-up e.g. monitoring programs.

The aim of this study is to demonstrate that it is not the variance of observed plant abundance data, but rather the way the plant abundance data are analysed, that is the problem. To address this problem, zero-inflated and over-dispersed stochastic models of single-species plant cover data from a hierarchical sampling procedure will be discussed, and relevant statistical distributions that describe the stochastic ecological processes underlying the observed variation will be suggested for different types of plant cover data. In another paper, it will be shown how the introduced distributions may be used to fit regression models of e.g. abiotic predictors to hierarchical plant cover data (Damgaard, 2008).

Standard statistical methods to deal with augmented variation of e.g. binomial data already exist, but since the augmented variance in a hierarchical sampling procedure may be caused by two completely different ecological processes operating on different spatial scales, the purpose of the suggested methods is to be able to partition the augmented variation into its two components in such a way that the underlying ecological processes may be studied more directly.

2. Materials and methods

2.1. Plant cover measurements

The most common way to measure plant cover in herbal plant communities, is to randomly throw a circle or a square onto the ground and then make a visual assessment of the relative area covered by the different species (Kent and Coker, 1992). The visually assessed cover of a plant species is then recorded as a continuous variable between 0 and 1, or divided into interval classes (e.g. Braun-Blanquet, 1964) as an ordinal variable. In visual assessments of plant cover, small plants tend to be overlooked, and it may be relevant to specify a minimum detection level of plant cover, i.e. to allow for the fact that a plant species may be present in a minute abundance even though the species has not been recorded at the plot. Such a minimum detection level depends on the investigated plant community, the size of the investigated plot, the amount of time allocated to observe a plot, and the person who makes the observations, but may typically be set to about 1%.

Although practical and fast, the visual assessment of cover has been criticised for being too subjective, i.e. too strongly dependent on the person who makes the observation (Floyd and Anderson, 1987; Kennedy and Addison, 1987). Consequently, an alternative more objective methodology, called the pin-point method (or point-intercept method), has been widely employed (Levy and Madden, 1933; Kent and Coker, 1992). In a pin-point analysis, a frame with a fixed grid pattern is placed randomly above the vegetation, and a pin is inserted vertically through one of the grid points into the vegetation. The different species touched by the pin are recorded at each insertion. The cover of plant species k in a plot, c_k, is now assumed to be proportional to the number of "hits" by the pin, $c_k = y_k/n$, where y_k is the number of pins that hit species k out of a total of n pins. Since a single pin in multi-species plant communities often will hit more than a single species, the sum of the plant cover of the different species may be larger than unity when estimated by the pin-point method. The sum of the estimated plant cover is expected to increase with the number of plant species in a plot and with increasing 3-dimensional structuring of the plants in the community.

In the above introduction to plant cover measurements, the focus has been on the practical methods of measuring plant cover: visual assessment and the pin-point method. However, for our purpose it is more appropriate to regard the two methods as the most common methodology for obtaining either continuously, ordinally or discretely distributed plant cover data. Recently, new methods of measuring plant cover are developed (e.g. Seefeldt and Booth, 2006), but they may all be classified according to whether they generate continuously, ordinally or discretely distributed plant cover data.

For all plant cover measurements that arise from a hierarchical sampling procedure, the data may be assumed to be adequately described by a two-stage stochastic process: i) A process that operates on a relatively large spatial scale, which is controlled by extinction events and colonisation events among sites. It is suggested that a zero-state process, from which only zero values are generated, may adequately model this process. ii) A process that operates on a relatively small spatial scale, which is controlled by seed dispersal plant growth, inter-specific competition, and environmental heterogeneity at the level of the site. This small-scale spatial process, which will generate overdispersed data compared to random expectations, depends on the method of measuring plant cover. The continuous plant cover data obtained by visual assessment are suggested to be adequately modelled by a beta distribution. Whereas the discrete plant cover data obtained by the pin-point method are suggested to be adequately modelled by a generalised binomial distribution (or Polya–Eggenberger distribution) (Qu et al., 1993).

2.2. Pin-point data: Zero-inflated generalised binomial distribution (ZIGBD)

The stochastic variable Y_k is the number of pins that, when inserted vertically through one of the n grid points into the vegetation, touch individuals of plant species k. The stochastic variable Y_k is assumed to be generated by a two stage stochastic processes: i) a zero-state process, from which only zero values are generated with probability p, ii) and a discrete binomial state process, from which all of the non-zero and some of the zero values are generated according to a generalised binomial distribution (or Polya–Eggenberger distribution) with probability parameter q and an intraclass correlation parameter, δ, which measure the correlation between the outcomes of successive Bernoulli trials (Qu et al., 1993). If the intraclass correlation parameter, which is bounded between $-\min(q/(n-1-q), (1-q)/(n-1))$ and 1, is different from zero, i.e. data are correlated, then the density of Y is,

$$f_Y(y; n, p, q, \delta) = \begin{cases} p + (1-p)\dfrac{\varphi\left(\dfrac{(1-q)(1-\delta)}{\delta}, n\right)}{\varphi\left(\dfrac{1}{\delta}-1, n\right)} & y = 0 \\ (1-p)\binom{n}{y}\dfrac{\varphi\left(q\left(\dfrac{1}{\delta}-1\right), y\right)\varphi\left(\dfrac{(1-q)(1-\delta)}{\delta}, n-y\right)}{\varphi\left(\dfrac{1}{\delta}-1, n\right)} & y = 1, 2..., n \end{cases} \quad (1)$$