Oecologia
DOI 10.1007/s00442-008-1017-y

COMMUNITY ECOLOGY - ORIGINAL PAPER

Beals smoothing revisited

Miquel De Cáceres · Pierre Legendre

Received: 6 November 2007 / Accepted: 20 February 2008

Abstract Beals smoothing is a multivariate transformation specially designed for species presence/absence community data containing noise and/or a lot of zeros. This transformation replaces the observed values of the target species by predictions of occurrence on the basis of its co-occurrences with the remaining species. In many applications, the transformed values are used as input for multivariate analyses. As Beals smoothing values provide a sense of "probability of occurrence", they have also been used for inference. However, this transformation can produce spurious results, and it must be used with caution. Here we study the statistical and ecological bases underlying the Beals smoothing function, and the factors that may affect the reliability of transformed values are explored using simulated data sets. Our simulations demonstrate that Beals predictions are unreliable for target species that are not related to the overall ecological structure. Furthermore, the presence of these "random" species may diminish the quality of Beals smoothing values for the remaining species. A statistical test is proposed to determine when observed values can be replaced with Beals smoothing predictions. Two real-data example applications are presented to illustrate the potentially false predictions of Beals smoothing and the necessary checking step performed by the new test.

Keywords Barro Colorado Island · Beals smoothing · Binary data · Community ecology · Randomization model

Introduction

Community composition data tables (i.e., species-plot data) are routinely used to study community structure and processes. In such tables, rows usually represent sampling units or sites, columns are species, and entries correspond to the contributions of the species to the sampled sites (e.g., presence/absence, biomass or density estimates). The research hypothesis is often that community data tables contain the responses in species composition to the environmental conditions prevailing at the sampled sites (in a broad sense, including the abiotic environmental factors as well as biological interactions and disturbances). However, a well-known issue of multivariate analyses is the loss of sensitivity of resemblance measures as the environmental distance between sampling units increases. This occurs because abundance values for species are used as a surrogate measure for habitat suitability, and the information on suitability is lost whenever the species is absent (McCune 1994). Beals (1984) referred to this problem as the "zero-truncation problem", which is essentially similar to the well-known "double-zero problem" (Legendre and Legendre 1998). In order to lessen the zero-truncation problem, Beals (1984) introduced a data transformation, which he called the "sociological favorability index". This

Communicated by Scott Collins.

Electronic supplementary material The online version of this article (doi:10.1007/s00442-008-1017-y) contains supplementary material, which is available to authorized users.

M. De Cáceres (✉) · P. Legendre
Département de Sciences Biologiques, Université de Montréal,
succursale Centre-ville, C.P. 6128, H3C 3J7 Montréal,
Québec, Canada
e-mail: miquel.de.caceres.ainsa@umontreal.ca;
mcaceres@ub.edu

M. De Cáceres
Departament d'Estadística, Universitat de Barcelona,
Avda. Diagonal 645, 08028 Barcelona, Spain

index assesses the "probability of occurrence" of a target species at a given site on the basis of its joint occurrences with the remaining species in the data table. A similar approach can be traced back to the works of Swan (1970) and Brisse et al. (1980). Ten years later, McCune (1994) studied this transformation and proposed a new name for it, *Beals smoothing*. Apart from avoiding the problem of many plots not actually sharing any species, McCune (1994) found Beals smoothing to be "particularly effective" on noisy data. Niche theory tells us that species are found at locations where they encounter appropriate living conditions (Hutchinson 1957). This does not mean, however, that species will always be found whenever suitable environmental conditions occur. For example, there may be historical or physical factors limiting the access of a species to an ecologically suitable habitat. Sampling errors, such as species misidentification or insufficient sampling effort, may also plague the data table. Clearly, one of the most important benefits of the Beals transformation is that it fills these niche "gaps" with species occurrence predictions, thus smoothing out the "ecological noise". When this "noise" is removed by the transformation, the multivariate structure of the data table is greatly simplified, leading to a substantial increase in the proportion of variance represented in non-canonical ordination plots (McCune 1994; Schnittler et al. 2006).

The popularity of this transformation in community ecology increased after its inclusion in statistical packages (McCune and Mefford 1999; Oksanen et al. 2008). In recent years, Beals smoothing has been used as a transformation for binary data prior to multivariate community analyses—mostly non-canonical methods, such as Bray–Curtis ordination (Ellyson and Sillett 2003), CA/DCA (Joy and Death 2000; Holz and Gradstein 2005), and nonmetric multidimensional scaling (NMDS; Kimball et al. 2004; Lee 2004; Marra and Edmonds 2005; North et al. 2005; Whitehouse and Bayley 2005; Beauchamp et al. 2006; Schnittler et al. 2006), and also clustering (Whitehouse and Bayley 2005). The communities under study are varied in the type of organisms, comprising vascular plants (Kimball et al. 2004, Lee 2004; North et al. 2005), lichens and bryophytes (Ellyson and Sillett 2003; Holz and Gradstein 2005; Whitehouse and Bayley 2005), arthropods (Marra and Edmonds 2005), fishes (Joy and Death 2000), and fungi (Beauchamp et al. 2006; Schnittler et al. 2006). Beals smoothing values have also been used for inference purposes because they provide a sort of "probabilities of occurrence" (Ewald 2002; Münzbergová and Herben 2004).

Despite its merits, it is important to state that replacing observed species values by Beals smoothed values in multivariate analyses must be done with caution for at least three reasons. First, it may produce erroneous or spurious results. Beals smoothing can produce consistent trends even from series of random numbers (McCune 1994; McCune and Grace 2002). Second, a species may have a higher probability of occurrence in a site where it does not occur than in sites where it occurs (Oksanen et al. 2008). Third, as this transformation highlights the "estimated" ecological niche for the target species, it can distort or mask spatial or temporal changes in community structure (Brodeur et al. 2005). Hence, the appropriateness of applying the Beals smoothing transformation also depends on the ecological question of interest. These may be the main reasons why this transformation is not widely used among ecologists. Clearly, it needs to be further studied before its use is promoted and extended.

A first purpose of this paper is to analyze the reliability of Beals smoothing in several situations comprising different data table sizes or incorporating noise in the data. Surprisingly, in many of the applications cited above, the presence of many zeros in the data table at hand was taken as a sufficient argument for using the transformation. In our opinion, one should instead check whether Beals smoothing predictions are expected to be reliable enough to replace the observed values or to make inferences. This would reduce the chances of obtaining spurious results. Therefore, the second purpose of this paper is to develop a statistical test to confirm or reject the Beals smoothing prediction for each target species. The structure of the paper is the following. First, the Beals smoothing function is presented, and the ecological bases underlying the transformation are stated. The next section proposes a statistical test whose null hypothesis non-rejection implies rejecting the Beals smoothing predictions for a given target species. This is done by measuring the amount of agreement between the predicted and observed values of the target species and by assessing whether a similar match could appear by chance. After that, the reliability of the Beals smoothing function and the statistical power of the proposed test are studied by extensive simulations. In the application section, we explore the potentially spurious effects of the Beals transformation on two forest data sets with contrasting ecological characteristics.

The Beals smoothing method

The Beals smoothing function

Let $\mathbf{X}$ be a community data table with r rows (i.e., sampling units) and p columns (i.e., species) containing the abundance values of a species. The Beals smoothing function considers only the species incidence information of $\mathbf{X}$, referred to as $\mathbf{X}^0$. Each $\mathbf{X}^0$ cell element, x^0_{ik}, contains either one or zero, indicating the corresponding presence or absence of species k in sampling unit i. The first step of the

Beals smoothing transformation consists in obtaining a symmetric matrix, $\mathbf{M}$, whose values are the number of joint occurrences for every pair of species (i.e., $\mathbf{M} = \mathbf{X}^{0t}\mathbf{X}^0$). The vector of its diagonal elements, $\mathbf{n} = \mathrm{Diag}(\mathbf{M})$, contains the number of occurrences of each species. After that, one can compute the Beals smoothing value (Beals 1984; McCune 1994), b_{ij}, which is the "probability" that a given target species j occurs in sampling unit i:

$$b_{ij} = (1/s_i)\sum_{k=1}^{p} \frac{m_{jk}x_{ik}^0}{n_k}, \qquad (1)$$

where s_i is the number of species in unit i, m_{jk} is the number of joint occurrences of species j and k, and n_k is the number of occurrences of species k. Despite its formal simplicity, some important issues about this transformation need to be remarked on in order to understand it properly:

1. The term m_{jk}/n_k is actually an estimate of the probability of occurrence of species j conditional to the known occurrence of species k. We will denote this estimate as $\hat{p}_{j/k}$. Under this interpretation, b_{ij} is an *average of estimated conditional probabilities*.
2. Highly frequent species tend to show high $\hat{p}_{j/k}$ values because many species often co-occur with highly frequent species. Conversely, those species with which a low-frequency (i.e., rare) target species jointly occurs are likely to occur in many other units where the target species is absent. Hence $\hat{p}_{j/k}$ is usually low for a rare target species. As shown in Electronic Supplementary Material (ESM)S1, if the occurrence of the target species j is independent of the occurrence of the remaining species, then the expected value of b_{ij} is simply the overall frequency of the target species.
3. In the original formulation (Beals 1984; McCune 1994), the target species was included in the summation. This causes a bias towards higher Beals values in sites where the target species is already present, which is not a nuisance when the objective is the replacement of the abundance values of the observed species, but it is a handicap when inference has to be done (Münzbergová and Herben 2004). To remove this bias, the target species must be excluded from the summation in Eq. (1).
4. The table of species joint occurrences may be obtained from a different source than $\mathbf{X}^0$, for instance, by means of a species-plot database or a bootstrapped sample of $\mathbf{X}^0$ rows (Münzbergová and Herben 2004)—provided the species indices match, nothing prevents $\mathbf{M}$, and therefore the $\hat{p}_{j/k}$ values, to be computed from a different reference table $\mathbf{Y}^0$ (r' x p).

In our opinion, the Beals smoothing function should be formally redefined to better reflect the above considerations. As stated in Beals (1984), this function provides an assessment of the "sociological favorability" of a species in a target sampling unit i. The parameters of the function are a vector of estimated conditional probabilities, $\hat{\mathbf{p}}_{j/}$, and the vector of species incidence values in the target sampling unit, $\mathbf{x}_i^0$:

$$b_{ij} = b_j(\hat{\mathbf{p}}_{j/}, \mathbf{x}_i^0) = \frac{\sum_{k=1,k\neq j}^{p} \hat{p}_{j/k}x_{ik}^0}{\sum_{k=1,k\neq j}^{p} x_{ik}^0}, \qquad (2)$$

where $\hat{\mathbf{p}}_{j/} = Diag(\mathbf{Y}^{0t}\mathbf{Y}^0)^{-1}\mathbf{Y}^{0t}\mathbf{y}_j^0$ is computed from a possibly distinct reference table $\mathbf{Y}^0$ ($r' \times p$). It is clear that, in order to compare Beals values corresponding to different sampling units, the vector of conditional probabilities $\hat{\mathbf{p}}_{j/}$ must always be the same. Note that when the species of interest is the only species observed in a sampling unit, Eq. (1) yields 1, while Eq. (2) yields 0. The former thus retains the observed value as the prediction, whereas the latter emphasizes the fact that there is a lack of co-occurrence information to make any prediction. Routinely, applications of the smoothing function may preferably be done with Eq. (1) because Eq. (2) is more prone to the problem pointed out in Oksanen et al. (2008).

Extending the Beals smoothing function to the case of abundance data is possible. We studied two possible generalizations of the Beals smoothing function which take into account abundance values of species in either the reference table $\mathbf{Y}$ or the target unit vector $\mathbf{x}_\mathrm{i}$, respectively. Since these are not considered in the following sections, they have been included in ESM S2 for the benefit of interested readers.

Ecological basis underlying the application of Beals smoothing

It is important to consider the ecological model underlying Beals smoothing before applying it to data at hand. This model essentially assumes that the pattern of occurrence of the target species can be predicted from its joint occurrences with the remaining species. If this is true, there should be an overall concordance between the observed and smoothed values of a target species, allowing an ecologist to replace the former by the latter. We believe this assumption can be verified in at least two different ecological situations:

1. *Environmental control*. If the observed occurrences of the non-target species follow their corresponding ecological niches, then the combination of a set of species from a sampling unit provides an integrated estimate of the prevailing environmental conditions across the sampled habitat. If the observed pattern of the target species is also environmentally controlled,

the combination of the species occurrences from a habitat with the data on species co-occurrences, carried out in Eq. (2), will provide a valid estimate of habitat suitability for the target species (Münzbergová and Herben 2004).

2. *Species associations*. Since the environmental conditions are not explicitly treated in Beals smoothing, there can be situations where there is high predictability of the target species from co-occurrences even though the species may not be environmentally controlled (i.e., if there are groups of correlated species). Perhaps this situation justifies the original name of the function: the "sociological favorability index" (Beals 1984).

In the first ecological situation (1), the observed values of a target species can be replaced by Beals predictions only if the observed pattern is mainly controlled by environmental factors and there is a sufficient proportion of the non-target species that is also environmentally controlled. Both hypotheses could be checked by means of multiple regression or canonical analysis using environmental variables as explanatory variables. In contrast, in the second situation (2), the replacement would not be valid for those target species that are not significantly associated (either positively or negatively) with any other species. Again, there are statistical methods to check this hypothesis (e.g., Legendre 2005). Nevertheless, in both ecological situations it is assumed that some kind of "ecological structure" exists (purely sociological and/or environmentally driven). The Beals smoothing approach for a target species is valid only if the species is related to that structure. We will hereafter use the words "ecological structure" to generally refer to any of these two ecological situations. Note that the presence of species not related to the structure (that is, species that are neither environmentally controlled in the same way as the others nor significantly associated), as well as other sources of "noise" in the reference table, can affect the reliability of Beals predictions. There may be cases where such structure is very weak, or even non-existent.

Selecting suitable species for Beals smoothing

One of our concerns here is to find a way to statistically test whether the observed values of the target species can be replaced by the Beals smoothing predictions. This turns out to be the same as testing the following ecological question: whether the target species is related to the ecological structure or not. Under the null hypothesis that negates this relationship, the joint occurrences with the non-target species will be at random, and the expected value of the Beals function will be the overall frequency of the target species; this statement is valid only when using Eq. (2). In this situation, the distribution of Beals smoothing values for sites where the species has been observed will be similar to the distribution for sites where it is absent, and both distributions will have the same mean. On the contrary, if the target species is related to, at least, some of the remaining species, then the Beals smoothing values in sites where these species occur will tend to be higher than the values in sites where they are absent. As a result, the two distributions of Beals values will become distinguishable. On the basis of this argument we propose the *Beals test*:

- H_0: The target species occurs randomly with respect to all the non-target species. The distribution of Beals values for sites where the target species occurs and that for sites where it is absent are undistinguishable.
- H_1: The target species is related to the ecological structure produced by, at least, some of the non-target species. The two distributions can be distinguished.

The degree of distinction between the two distributions can be measured with several statistics accounting for the variance in the distributions (e.g., Mann–Whitney's U or a t statistic). However, as one or both distributions may contain few values, their variance may not be estimated properly, and a simpler statistic comparing central positions may be more powerful. After some preliminary investigations, we decided to use the difference between medians, $S = \mathrm{Med}(A) - \mathrm{Med}(B)$, where A is the set of Beals values in sites where the target species occurs, and B is the set of values in sites where it is absent. The statistic S observed for a given target species (S_{obs}) has to be compared with a reference distribution, which can be generated using randomization methods. In order to devise a proper randomization test, we have to take into account the fact that at least some of the non-target species may constitute an ecological structure, which should not be altered. Following the null hypothesis, a suitable randomization consists in permuting the column values of the target species while keeping the other columns of the table intact. This implements a null model for the joint occurrence, or association, of the target species and the remaining species (Gotelli 2000; Legendre 2005). An unrestricted permutation model may, however, generate situations that are ecologically unreliable if the beta-diversity of the data is very high, which increases the likelihood that the occurrences of the species will be placed in sampling units of very different ecological nature. Therefore, some kind of environmental constraint in the null model seems advisable (e.g., Peres-Neto et al. 2001). We decided to restrict permutations to among the sampling units corresponding to non-zero Beals smoothing values; this is equivalent to restricting the permutations to the sites occupied by any of the species co-occurring with the target species (but see the Application section for another type of restriction). Following each randomization step, the test statistic is re-computed and stored in S_{rnd}. After all permutations are carried out, the observed test statistic S_{obs} is compared to the distribution of random values generated under the null hypothesis, and the probability of the data under H_0 is calculated as $p = (1 + \text{no. } S_{\text{rnd}}$ values equal to or larger than $S_{\text{obs}})/(\text{no randomizations} + 1)$. The "1" in the numerator adds the observed value to the distribution, as recommended by Hope (1968). If one is interested in preserving an experiment-wise type I error rate, the p values may be adjusted for multiple testing (Sidak 1967; Holm's 1979).

Simulation methods

The objectives of the present simulation study were twofold. First, we were concerned with assessing the reliability of Beals smoothing predictions for various numbers of sampling units or species in the reference table and in the presence of noise in the data. A second aim was to evaluate the type I error rate of the proposed statistical test, and its power, under the same experimental conditions.

Simulated data

Simulated community data tables were created using the program JCOMPAS, which is an adaptation of Minchin's (1987) COMPAS that forms part of the program GINKGO (Bouxin 2005; De Caceres et al. 2007). In the COMPAS model, multivariate species abundances are simulated in an ecological space using unimodal beta functions (Austin 1976). Since the Beals smoothing transform is computed without reference to a classification and presupposed gradients (Ewald 2002), we considered the nature of the modeled ecological structure to be of secondary importance. For simplicity, our simulated ecological space consisted of a single ecological gradient in the interval [0–100]. Modal coordinates were chosen uniformly along and beyond the gradient (from −150 to 150 units to avoid border effects), and modal abundances were constantly 100 for all species. The parameters related to the form of the species distributions were sampled from a uniform distribution in the interval [0.3–0.5]. The species responses were initially generated with neither systematic trend nor qualitative or quantitative noise (Minchin 1987). Therefore, abundance values of species were directly related to habitat suitability, as specified in the model.

We generated two initial data sets, each containing 1000 independent species and 1000 sampling units uniformly distributed along the simulated ecological gradient (step 1, Fig. 1). The first data table was obtained by sampling species niche breadth values from a normal distribution with a mean of 50 units and standard deviation of 10 units; beta-diversity was approximately two half-changes (=100/50). The second data table had higher beta-diversity (four half-changes) since the species niche breadth values were sampled from a normal distribution with a mean of 25 units and standard deviation of 5 units. These two data sets were then re-sampled in order to generate smaller subsets of 100 distinct size combinations (from 10 species × 10 sites to 100 species × 100 sites). Sites were sampled with replacement, but not species. Each of the resulting data tables was duplicated to distinguish between the target and reference tables (**X** and **Y**, respectively, step 2, Fig. 1).

Noise addition

Data tables **X** and **Y** were then subjected to sources of noise (step 3, Fig. 1). We will denote with $\mathbf{X}'$ and $\mathbf{Y}'$, respectively, the data tables resulting from noise addition. Three kinds of noise were considered in our simulation study:

A. "Random species" were generated by independently permuting the column values corresponding to a fraction of the species. The same permutation was applied to both **X** and **Y**.
B. "Two ecological structures" were represented by two uncorrelated groups of species. We divided the species

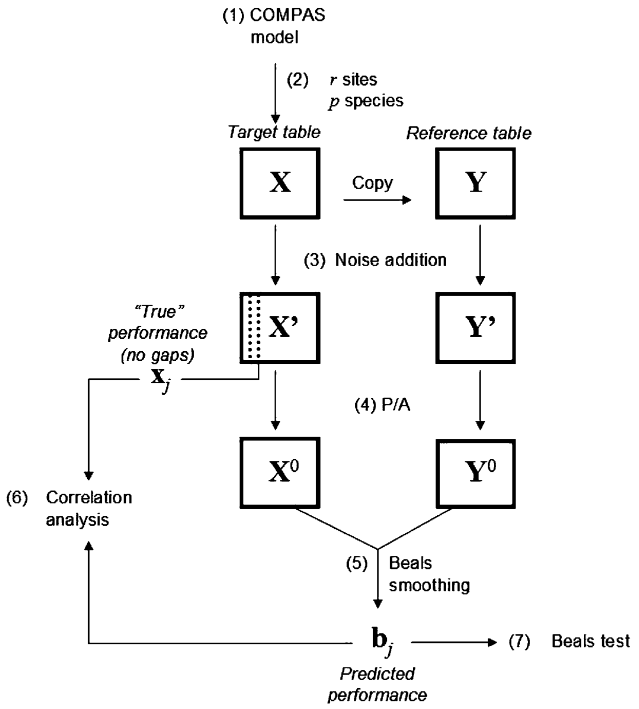

Fig. 1 Schematic representation of the steps performed in the simulation study (see text for details)

into two equally sized groups; entire rows were permuted within the sub-table corresponding to one of the two groups. Again, the same permutation was applied to both $\mathbf{X}$ and $\mathbf{Y}$.

C. "Gap noise" was generated by randomly introducing zeros (i.e., "gaps") in the modeled species abundances. In this way, we simulated the absence of species due to reasons not related to its ecological niche. Each gap followed a Bernoulli distribution. That is, we fixed a probability of occurrence p and generated vectors of [0/1] random values that were used as multipliers of the generated species abundances; an abundance was 0 if the corresponding Bernoulli value was 0. Only the reference data table $\mathbf{Y}'$ contained gap noise, and not the target table $\mathbf{X}'$; since this would be equivalent to considering fewer species in sampling units, and we wanted to separate the gap noise error effect on $\hat{\mathbf{p}}_{j/}$ from the effect of species richness (s_i).

Eight distinct error scenarios were studied:

1. Noise free data: here data tables $\mathbf{X}'$ and $\mathbf{Y}'$ were exactly the same as $\mathbf{X}$ and $\mathbf{Y}$.
2. Random (i.e., permuted) target species: values of the target species were permuted (noise type A). In this scenario, an ecological structure exists in the remaining set of species but the target species does not belong to it.
3. 50% random (i.e., independently permuted) non-target species: half of the non-target species were permuted (noise type A). They constituted a source of error for predicting the target species.
4. 50% of species randomized in block: half of species were permuted in block (noise type B). In this scenario, those species belonging to one group constituted a noise source for the prediction of target species belonging to the other.
5. 100% (i.e., independently permuted) random species: all species were permuted independently (noise type A). In this scenario, there was no ecological structure at all.
6. 50% of gap noise added to the non-target species only: this scenario tried to simulate a partial decrease in predictive power of all the non-target species. For each target species, gaps (noise type C) were added to the remaining columns of the reference table $\mathbf{Y}'$.
7. 50% of gap noise added to the target species only: this scenario simulated a partial violation of the target species relationship to the ecological structure. For each target species, gaps (noise type C) were added to its corresponding column in the reference table $\mathbf{Y}'$.
8. 50% of gap noise added to all species: we combined here the error effects from scenarios (6) and (7), i.e., gap noise for both the target and non-target species. This was perhaps the most realistic scenario among all.

Statistical analyses

Under all simulated conditions (i.e., under all error scenarios and for all numbers of species and numbers of sampling units), we did the following for each target species. First, we took the vector of target species abundance values in $\mathbf{X}'$, denoted $\mathbf{x}_j$ in Fig. 1. Note that this vector never contained gap noise, but its values could have been permuted. Permuted species are those whose ecological performance is unaffected by the simulated ecological gradient. Therefore, vector $\mathbf{x}_j$ can be considered to contain the "true" ecological performance of the target species. Second, we transformed data tables $\mathbf{X}'$ and $\mathbf{Y}'$ into their corresponding incidence data tables $\mathbf{X}^0$ and $\mathbf{Y}^0$ (step 4, Fig. 1). Beals smoothing values were then computed using Eq. (2) (step 5, Fig. 1). As this equation does not take into account the observed presence or absence of the species in the target site, it enables us to know how effective the Beals function is in filling species "gaps". We assessed the reliability of the Beals smoothing values by calculating the Pearson correlation coefficient between the target species' predicted values, $\mathbf{b}_j$, and the "true" abundance values in $\mathbf{x}_j$ (step 6, Fig. 1). Due to the zero truncation, zero-simulated abundance values are not an indication of species suitability. Therefore, sampling units where the modeled abundance was zero were excluded from the correlation analysis. In contrast, those sampling units containing zeros due to gap noise were included in the comparison (note that $\mathbf{X}'$ did not contain gap noise). We also restricted the correlation analysis to those target species with positive-simulated abundances in at least five sampling units; this insured that it was possible for the computed correlation coefficients to reach the 1% significance level. The Beals test was run on the same target species included in the correlation analysis (step 7, Fig. 1). The number of permutations was set to 199 and the significance level to $\alpha = 0.05$.

Steps 2–7 were repeated as many times as was necessary in order to have at least 1000 target species. We then averaged the corresponding 1000 correlation values to obtain an average correlation assessing the reliability of Beals predictions under the simulated conditions. The statistical power of the Beals test was assessed by counting the proportion of target species for which the test rejected the null hypothesis and dividing this value by the number of tests. Type I error rate was also assessed by modeling incidence data using the Bernouilli distribution, where the probability for each species was equal to its corresponding relative frequency in $\mathbf{X}$. Data re-sampling, noise addition, and statistical analyses were performed using functions written in the *R* statistical language (R Development Core Team 2007) and are available under request.

Simulation results

Reliability of Beals smoothing values

Let us first consider the results on the error-free condition (scenario 1). Figure 2a and b shows the average Pearson correlation between predicted performance and true simulated performance that was computed in data sets of different sizes. The two axes of the representation can be interpreted to be independent sources of information. The number of sampling units indicates the size of the database used to compute conditional probabilities, $\hat{\mathbf{p}}_{j/}$. Alternatively, the higher the number of species in the data table, the higher the species richness in the vector of sampling units $\mathbf{x}_l^0$. As expected, an increase in either the number of sampling units or species richness had a positive effect on the reliability of Beals smoothing as a predictor of species performance. The differences between Fig. 2a and b are due to the following. For the same data set size, high beta-diversity data (right) had fewer joint occurrences in the reference data table than low beta-diversity data (left), so more sites were needed to obtain the same quality of conditioned probability estimations. At the same time, in high beta-diversity data, the sampling units had lower species richness and, consequently, predictions were less reliable than in low diversity data for the same total number of species.

We observed distinct effects on the reliability of Beals smoothing values in low diversity data depending on the noise scenario. Under permutation of the target species values (scenario 2) the predicted values were totally uncorrelated to the modeled abundances in all cases. Random permutation of 50% of the species (scenario 3) is equivalent to including some species with low predictive power; as a result, more species were needed to ensure a proper representation of the ecological structure (Fig. 3a) than with the noise-free condition (Fig. 2a). Block randomization (scenario 4) differed from the previous scenario in that species were not independently randomized. Consequently, the disturbance on the predictions for target species belonging to the other block was not cancelled out, and the overall noise effect was somewhat stronger (Fig. 3b). Under the permutation of all species (scenario 5), the results were the same as under the permutation of the target species (scenario 2): correlation was non-existent in all cases. The addition of 50% gaps into the non-target species of the reference table **Y** (scenario 6) had almost no effect on the reliability of Beals smoothing predictions (Fig. 3c). Whereas gap noise in non-target species *k* diminished the amount of sampling units used to compute the conditional probability, the target species *j* still occurred roughly in the same proportion of sampling units where *k* was found. In contrast, when gaps were present in the target species (scenario 7), the reliability of Beals smoothing predictions was substantially lower for a given number of sampling units (Fig. 3d). Our explanation in this case is that the gaps in target species *j* yielded an underestimation of $p_{j/k}$ for all non-target species *k* that were ecologically close to the target species. Consequently, their estimated conditional probability values became similar to the values corresponding to species that were less ecologically related to the target species. In short, the $\hat{p}_{j/k}$ values lose their predictive power. The addition of more species did not improve the situation because the gaps were in the target species and not in the non-target species. Only a larger data table (i.e., more sampling units) overcame this problem; then the number of occurrences of the target species became sufficiently large to ensure the quality of the $\hat{\mathbf{p}}_{j/}$ estimates. Finally, when gap noise was present in all species (scenario 8), the results were almost equal to those of scenario 7. Similar noise effects were observed for the high-diversity data sets (results not shown). We repeated the same study using Spearman instead of Pearson correlation, and the qualitative interpretation of results was the same.

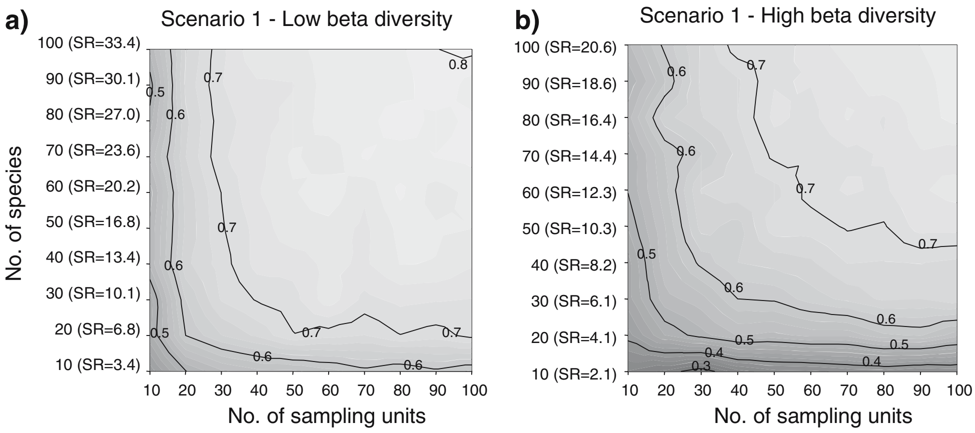

Fig. 2 Average correlations between the simulated abundance values and the Beals smoothing values (computed from presence–absence data) for different sizes of data tables and without noise (scenario 1). *SR* Mean species richness per sampling unit in the target data tables

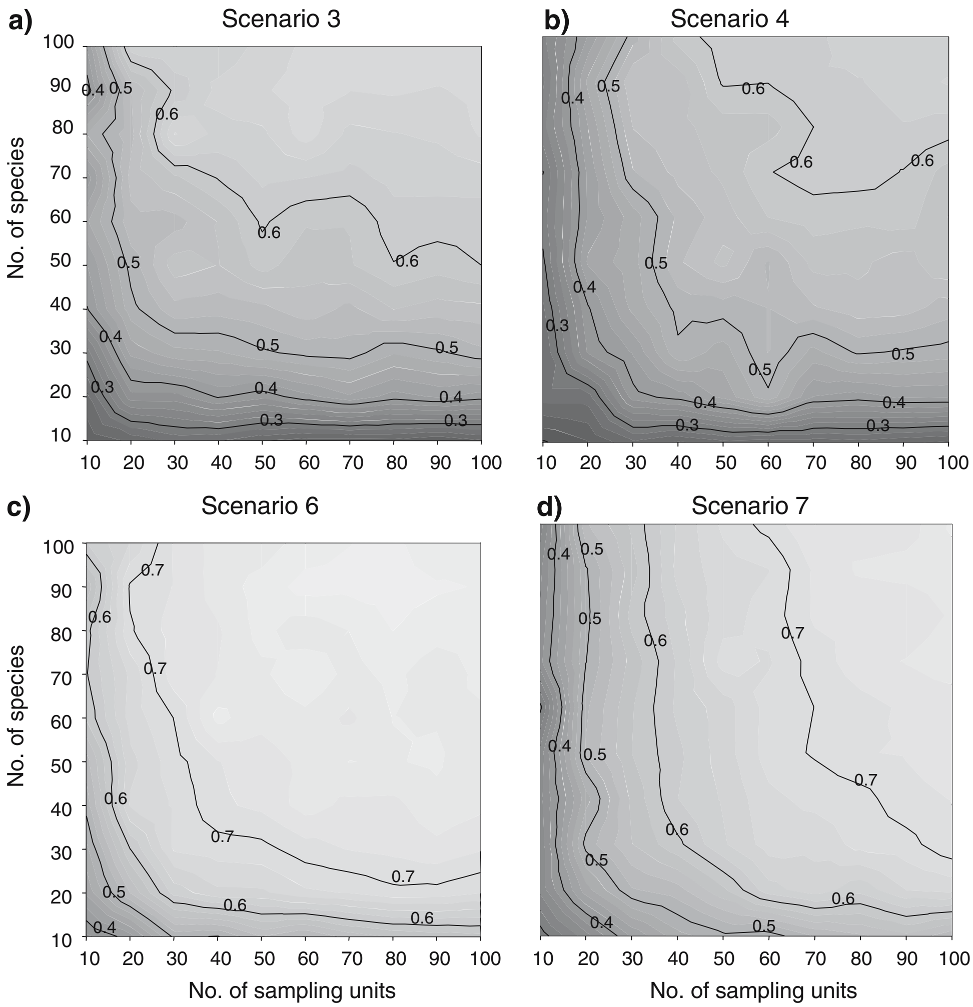

Fig. 3 Average correlations for low-diversity data table sizes and different noise conditions. **a** Scenario 3: half of the non-target species values permuted independently, **b** scenario 4, data rows permuted within a block comprising half of the species, **c** scenario 6, 50% of gaps added to all non-target species of the reference table, **d** scenario 7, 50% gaps added to the target species of the reference table. Mean species richness per sampling unit in the target data tables is the same as in Fig. 2a

Statistical power and type I error rate of the Beals test

We used the rate of rejection of the null hypothesis on data simulated using the Bernouilli distribution to assess the type I error rate of the test. The empirical type I error rate was slightly under the 5% significance level (its 95% confidence interval was [0.0427, 0.0456]) consequently, the permutation test cannot be considered exact, although it is still valid. We were also interested in knowing whether or not the Beals test was capable of making the distinction between "random" and "true" species, and how its statistical power was affected by the amount of information and the various noise situations. Figure 4a and b show the rates of rejection of the null hypothesis for data tables of distinct sizes and beta-diversity values and no noise (scenario 1). The number of species in the data tables had little impact on statistical power—very few related species are necessary to display a sociological structure—which mostly depended on the number of sampling units. With a minimum number of sampling units—say $r = 20$–30 for both high and low beta-diversity data—the power of the test was above 95%.

The power results of the Beals test under the noisy scenarios and low-diversity data were the following. As could be expected, permuting the target species (scenario 2) resulted in it being significantly related to the structure 5% of the time. By contrast, the Beals test was quite robust (i.e., its power was quite stable) in the presence of non-target permuted species (scenario 3); that is, the presence of 50% permuted species did not severely affect the probability of rejecting the null hypothesis for non-permuted target species (Fig. 5a). When species were divided into two blocks (scenario 4), all species had to be detected as informative since they were either related to one ecological structure or the other. As can be seen in Fig. 5b, the presence of two independent ecological structures only slightly diminished the power of the method, since within each structure every target species had its ecological niche. Under the permutation of all species (scenario 5), the ecological structure was completely lacking, and the rate of rejection was again near 5%. Under the addition of 50% gap noise to the non-target species of the reference table (scenario 6, Fig. 5c), the test was almost as powerful

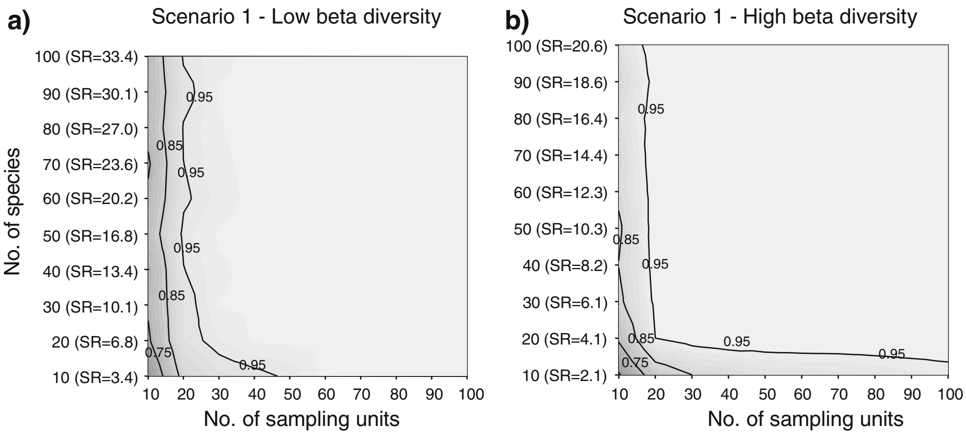

Fig. 4 Statistical power of the Beals test for different data table sizes and without noise (scenario 1). *SR* Mean species richness per sampling unit in the target data tables

(lower) as for the noise-free condition. In contrast, the presence of gap noise in the target species (scenario 7) severely affected the statistical power of the Beals test (Fig. 5d), and far more sampling units were needed to attain the rate of 95% of true species recognition than in noise-free data. Our explanation here is that the more niche "gaps" there are in the target species observed values, the higher the probability of confounding it with a true "random" species. Scenario 8 results were again similar to those of scenario 7.

Real-data applications

One of the conclusions from the previous simulation study is that Beals smoothing produces spurious predictions whenever the target species is not related to the ecological structure or whenever the ecological structure does not exist. We believe this fact fully justifies the application of the proposed test before the replacement of observed values by predicted ones. In this section we show with real data how the apparently outstanding results of the Beals transformation may be spurious on occasion. We studied two distinct forest data sets:

1. The Barro Colorado Island (BCI) data from the first tree census conducted on the 50-ha forest dynamics plot in 1981–1983 (Hubbell et al. 2005). We calculated the incidence of each tree species on grids of cells of five distinct sizes, ranging from 10 × 10 m (5000 cells of 100 m^2) to 100 × 100 m (50 cells of 1 ha). For each cell size we hence obtained a distinct data table.
2. The vegetation data from Bryce Canyon National Park (Utah, USA) (Roberts et al. 1988), part of the 'labdsv' library (Roberts 2006) of the *R* statistical language. This data table contains abundance values for 169 vascular plants identified at 160 sampling sites. Again, we only considered the species presence/absence information.

We started our analyses by computing the average, maximum, and minimum number of species per site (Table 1). As the BCI plot is part of a tropical forest, species richness values are much higher there than in Bryce Canyon, which corresponds to temperate vegetation. The species richness of the BCI cell is also higher for broad scales (larger cell sizes). This average number of species a priori ensures the reliability of the Beals transformation in most cells/sites. For large cell sizes, many tree species in BCI may occur in all cells. As these species are uninformative in terms of co-occurrences, we counted the number of non-trivial species, i.e., species which are not always present. If ecological structures are present, the number of sampling units is high enough in all data tables to provide high-quality values of conditioned probabilities.

In most applications, Beals smoothing is used as a method to filter out noise before ordination analyses. We ran principal component analysis (PCA) on the original presence–absence tables, and we noted in Table 1 the overall data variability and percentages of explained variation corresponding to the first two PCA axes. Due to the more complex structure of the tropical forest, the first two PCA axes recovered a smaller percentage of the information for BCI than for Bryce Canyon. The percentage of information recovered was also lower for fine-scaled data (10 × 10 m cells) than for broad-scaled data. We again ran PCA on the data tables resulting from the application of Beals smoothing (without discarding any species). Beals smoothing greatly simplified the data because the percentages of structure explained in two first PCA dimensions were much higher, whereas the total number of dimensions needed to fully represent the data was the same. Up to this point, one may ask whether this astonishing and apparently successful clarification of data was spurious. In order to answer this question, we ran the Beals

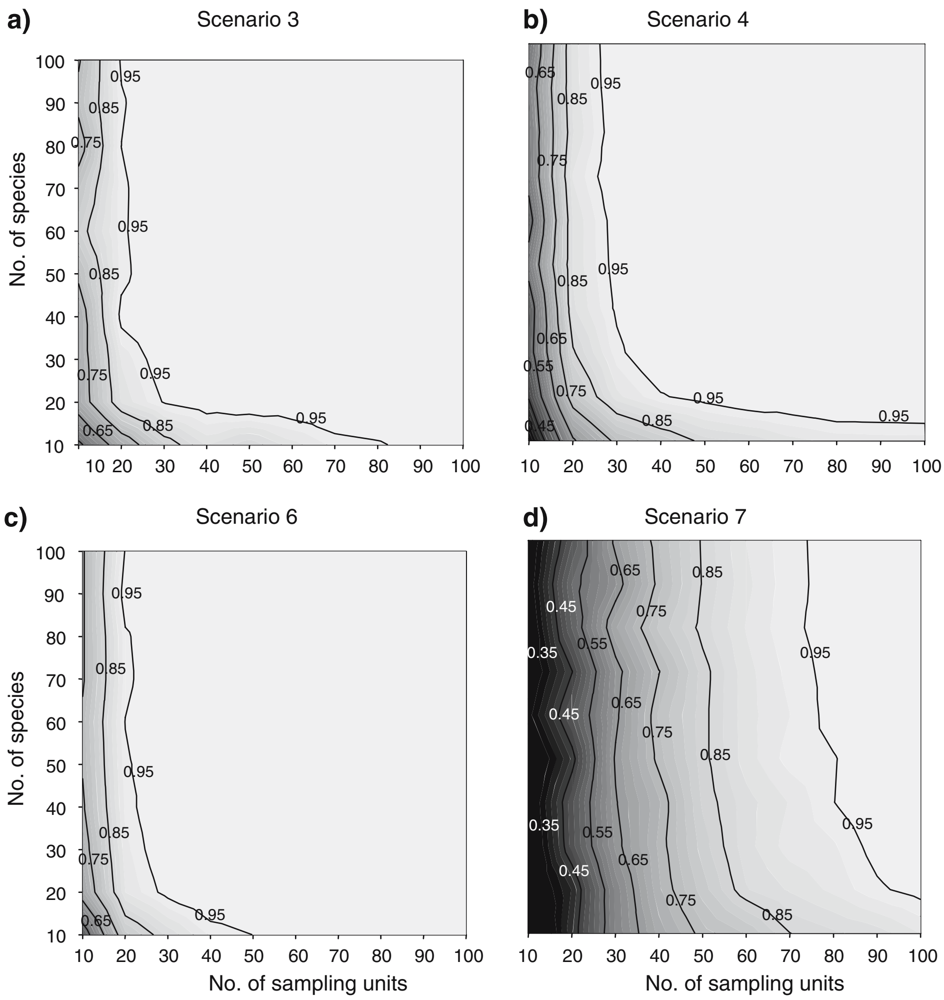

Fig. 5 Statistical power of the Beals test for different low-diversity data tables and different noise conditions. **a** 50% of the species values were permuted independently, **b** data rows were permuted within a block comprising 50% of the species, **c** 50% gaps were added to all non-target species of the reference table, **d** 50% gaps were added to the target species of the reference table. Mean species richness per sampling unit in the target data tables is the same as in Fig. 4a

test on all of the non-trivial species. The permutation approach described above is, however, hampered by the presence of autocorrelation in the distribution of BCI species. To correct for this autocorrelation effect, we restricted the permutations to those provided by a toroidal shift (Harms et al. 2001; Fortin and Dale 2005). After 199 permutations and using a significance level $\alpha = 0.05$, almost 60% of species were significant in the Bryce Canyon data set. Many of the non-significant species were also rare. In the case of BCI, the proportion of significant species was not the same for all cell sizes; the percentages were around 65% for the 10 × 10-m cells and decreased for broader scales, down to 12.9% of the non-trivial species for the 100 × 100-m cells. We studied whether ordination diagrams could be affected by the lack of reliability of Beals smoothing by computing the percentage of total variation corresponding to the set of significant species in both the original and transformed data tables [columns noted as ‘Significance (%)’ in Table 1]. For the Bryce Canyon data, the percentages were 85% for the original untransformed data table and 97% for the Beals-transformed. Thus, the 40% of “noisy” species (i.e., those that did not pass the Beals test) only accounted for 15% (i.e., 100–85) of the original variability and 3% (i.e., 100–97) of the transformed one. The latter percentage indicates that the interpretation of PCA ordinations on the transformed data is safe, since the information shown is mostly related to those species for which the transformation is valid. For the BCI data sets, the situation depended on the scale. In the finest scaled data set (10 × 10 m) more than half the species were significantly associated to the ecological structure, accounting for 90.8% of the original data variation and 95.8% of the transformed data variation. In this situation, ordination diagrams contain a substantial amount of information concerning significant species. However, the larger the cell size, the smaller the number of species that are significant and the lower the fraction of variability that they account for. In the worst case (100 × 100 m cells), in fact, 74% (i.e., 100–26) of the post-transformation variability corresponded to spurious predictions of the smoothing function. It is obvious that the transformation and the subsequent ordination in this case should not be used.

Discussion

Clearly, the Beals smoothing function needed to be further studied before recommendations could be made as to its use. In extending the work of McCune (1994), one of the aims of the present paper was to study the reliability of Beals smoothing values under different conditions. Assuming that an "ecological structure" exists and that the target species is related to it, the reliability of Beals smoothing depends on both the reliability of the estimated conditional probabilities and the species richness in the target sampling units. In turn, the reliability of the conditional probability estimations depends on the size of the reference data table. Münzbergová and Herben (2004) concluded their study by saying that "…any reasonably large database can be used". We attempted to go further and establish a practical threshold. Our simulations indicate that, in the absence of noise, a minimum of 40 sampling units are necessary in order to obtain a good correlation between the predicted values and the modeled species performance. The use of Beals smoothing with fewer sampling units may produce unreliable estimates. In terms of the number of species used to compute the function, we observed a rather strong robustness of Beals smoothing values to small species richness in the target sampling unit. For practical issues, we estimate that a minimum of ten species is necessary in the target sampling unit to ensure the quality of the predictions. While this may seem a rather liberal threshold, there are published applications of Beals smoothing to data with a richness of between five and ten species (e.g., Ellyson and Sillett 2003; Marra and Edmonds 2005; Beauchamp et al. 2006), and even below five species (e.g., North et al. 2005).

The above thresholds were given after naively assuming that the ecological structure was without noise and the data table had low beta-diversity (i.e., from Fig. 2a). In real applications, however, the predictive power of Beals smoothing is strongly affected by high beta-diversity values and/or the presence of noise in the species data. Specifically, species "gaps" due to spatial or temporal dynamics are expected to be a common source of noise in real vegetation data tables. Fortunately, the effect of these "gaps" on Beals predictions can be overcome if a sufficient number of sampling units are used in the reference table. This was probably the case of the BCI 10 × 10-m table, where the fine-scaled local spatial variability was counterbalanced by a large number of cells. In contrast, our results with simulated data indicate that the presence of

Table 1 Beals analysis results for Barro Colorado Island (BCI) and Bryce Canyon data. BCI tree data were analyzed at different cell sizes ranging from 10 × 10 m to 100 × 100 m

Dataset #	Cells/sites	Site species richness			Non-trivial	Beals significant		Original binary data			Beals-transformed data		
		Average	Minimum	Maximum	Number of species	Number of species	Species (%)	Total variation	PCA (%)	Significance (%)	Total variance	PCA (%)	Significance (%)
BCI 10 × 10	5,000	23	2	48	307	198	64.5	15.5	6.6	90.8	0.013	56.4	95.8
BCI 20 × 20	1,250	54	19	85	307	165	53.7	27.5	8.3	72.8	0.015	68.5	85.8
BCI 25 × 25	800	68	33	106	305	163	53.4	30.7	8.9	68.0	0.015	68.9	83.7
BCI 50 × 50	200	121	90	152	300	99	33.0	35.1	11.6	43.3	0.016	70.1	62.3
BCI 100 × 100	50	177	151	196	241	31	12.9	31.6	16.1	16.8	0.018	61.4	26.0
Bryce Canyon	160	14	3	27	169	100	59.2	10.8	28.0	85.0	1.629	82.3	96.9

PCA, Principal component analysis

The number of cells or sites is indicated in Cells/sites. The minimum, maximum and average species richness per cell/site, respectively. Non-trivial species are those whose pattern is neither a vector of zeros nor a vector of ones. Among the non-trivial species, we indicate the number and percentage of species significantly associated to the ecological structure following the Beals test (199 perm., $\alpha = 0.05$). We then show the comparison of original binary data and Beals-transformed data, in terms of total variation, the percentage of that variation that is represented on the first two PCA axes, and the percentage that is accounted for by the species that are significant following Beals test species not related to the sampled ecological structure calls for higher species richness in the target sampling units in order to achieve valid Beals smoothing predictions (Fig. 3a). Whereas sampling effort may be increased, it is not possible to sample more species once all existing species in the community have been found. “Random” species are likely to occur in real data. For example, community data can contain a subset of species whose pattern is related to environmental gradients that act at a scale different from the observation scale (i.e., microhabitat conditions may exist) or is sensitive to a different ecological factor. Furthermore, the species whose patterns of occurrence contain many “niche gaps” may eventually behave like “random” species. In sites with small species richness, the presence of those “random species” causes the Beals smoothing values to be insufficiently reliable, regardless of the number of sites. Reliability problems will also appear when using a reference data table **Y** where the species have different co-occurrence structures than those in the target data table **X** (for example, due to a different geographical extent). Summing up, it is clear that in many real situations the Beals transformation should be avoided. Users must use this technique with caution.

The statistical test presented in this paper offers several advantages. First, it is expected to give non-significant results whenever there is no structure at all or the target species is not related to it. Second, the power of the Beals test does not appear to be severely affected by low species richness, provided there is a minimum number of sampling units (95% power is achieved between 20 and 30 sampling units in Fig. 4a–b) and enough occurrences of the target species to allow a valid randomization approach. Third, the Beals test is also quite robust to the presence of random non-target species or to gaps in their distribution (Fig. 5a, c). We showed in the last section how this test can be used to prevent spurious applications of the transformation. We believe the Beals test can also be useful to community ecologists to select species prior to multivariate analyses where either (1) species in the community data table are expected to be an expression of the prevailing habitat conditions, or (2) species associations are important.

Even though the test proposed here is a safeguard from spurious results of the smoothing function, it does not guarantee its appropriateness. When using the Beals smoothing, we are imposing an ecological model onto our data and throwing out the variability that does not fit this model. If the model is consistent with the assumptions and objectives of subsequent analyses, one will generally obtain better results on the transformed data. For example, Beals smoothing (or the selection of species based on the Beals test) may improve the results of analyses aiming at the elucidation of the relationships between communities in terms of their “potential” composition, such as in unconstrained ordination or clustering. When explaining community data by taking into account environmental variables, smoothed data will usually give the appearance of improving the species–environment relationships because all variability not related to the environmental control hypothesis will likely have been filtered out. Statistical tests comparing the amount of signal versus noise (e.g., *F*-like tests) should be more powerful with filtered data. While those analyses may be valid statistically, special care has to be taken when interpreting results ecologically as they will be referring to the Beals model and not to the original data. To be cautious, we recommend users perform statistical inference on the original data, unless otherwise fully justified. The analyses whose objectives are incompatible with the Beals’ model must be avoided—for example, quantifying and/or testing the significance of the spatial or temporal variation of ecological communities.

Acknowledgments This work benefitted from comments by Pedro-Peres Neto on randomization methods and by Daniel Borcard and Artur Lluent on the ecological interpretation of the Beals smoothing function. The authors are especially grateful to Jari Oksanen, who suggested interesting real-data applications and provided several suggestions, and to Bruce McCune and David Roberts for their comments on previous versions of the manuscript. This research was funded by NSERC grant no. OGP0007738 to P. Legendre. The BCI forest dynamics research project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots. All experiments comply with the current laws of the country in which the experiments were performed.

References

Austin MP (1976) On non-linear species response models in ordination. Vegetatio 33:33–41

Beals EW (1984) Bray–Curtis ordination: an effective strategy for analysis of multivariate ecological data. Adv Ecol Res 14:1–55

Beauchamp VB, Stromberg JC, Stutz JC (2006) Arbuscular mycorrhizal fungi associated with *Populus–Salix* stands in a semiarid riparian ecosystem. New Phytol 170:369

Bouxin G (2005) GINKGO, a multivariate analysis package. J Veg Sci 16:355–359

Brisse H, Grandjouan G, Hoff M, de Ruffray P (1980) Utilisation d’un critère statistique de l’écologie en phytosociologie–exemple des forêts alluviales en Alsace. Coll Phytosociol 9:543–590

Brodeur RD, Fisher JP, Emmett RL, Morgan CA, Casillas E (2005) Species composition and community structure of pelagic nekton off Oregon and Washington under variable oceanographic conditions. Mar Ecol Prog Ser 298:41–57

De Cáceres M, Oliva F, Font X, Vives S (2007) GINKGO, a program for non-standard multivariate fuzzy analysis. Adv Fuzzy Sets Syst 2:41–56

Ellyson WJT, Sillett SC (2003) Epiphyte Communities on Sitka Spruce in an old-growth Redwood Forest. Bryologist 106:197–211

Ewald J (2002) A probabilistic approach to estimating species pools from large compositional matrices. J Veg Sci 13:191–198

Fortin MJ, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge

Gotelli NJ (2000) Null model analysis of species co-occurrence patterns. Ecology 81:2606–2621

Harms KE, Condit R, Hubbell SP, Foster RB (2001) Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. J Ecol 89:947–959

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:1979

Holz I, Gradstein RS (2005) Cryptogamic epiphytes in primary and recovering upper montane oak forests of Costa Rica–species richness, community composition and ecology. Plant Ecol 178:89–109

Hope ACA (1968) A simplified Monte Carlo test procedure. J R Stat Soc B 50:35–45

Hubbell SP, Condit R, Foster RB (2005) Barro colorado forest census plot data. Available at http://ctfs.si/edu/datasets/bci

Hutchinson GE (1957) Concluding remarks. Cold Spring Harb Symp Quant Biol 22:415–427

Joy MK, Death RG (2000) Development and application of a predictive model of riverine fish community assemblages in the Taranaki region of the North Island, New Zealand. NZ J Mar Freshw Res 34:241–252

Kimball S, Wilson P, Crowther J (2004) Local ecology and geographic ranges of plants in the Bishop Creek watershed of the eastern Sierra Nevada, California, USA. J Biogeogr 31:1637–1657

Lee P (2004) The impact of burn intensity from wildfires on seed and vegetative banks, and emergent understory in aspen-dominated boreal forests. Can J Bot/Rev Can Bot 82:1468–1480

Legendre P (2005) Species associations: the Kendall coefficient of concordance revisited. J Agric Biol Environ Stat 10:226–245

Legendre P, Legendre L (1998) Numerical ecology, 2nd English edn. Elsevier, Amsterdam

Marra JL, Edmonds RL (2005) Soil arthropod responses to different patch types in a mixed-conifer forest of the Sierra Nevada. For Sci 51:255

McCune B (1994) Improving community analysis with the Beals smoothing function. Ecoscience 1:82–86

McCune B, Grace JB (2002) Analysis of ecological communities. MjM Software Design, Gleneden Beach

McCune B, Mefford MJ (1999) PC-ORD. Multivariate analysis of ecological data, Version 4. MjM Software Design, Gleneden Beach

Minchin PR (1987) Simulation of multidimensional community patterns towards a comprehensive model. Vegetatio 71:145–156

Münzbergová Z, Herben T (2004) Identification of suitable unoccupied habitats in metapopulation studies using co-occurrence of species. Oikos 105:408–414

North M, Oakley B, Fiegener R, Gray A, Barbour M (2005) Influence of light and soil moisture on Sierran mixed-conifer understory communities. Plant Ecol 177:13–24

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Stevens MHH (2008) Vegan: community ecology package. R package version 1.11-0. http://cran.r-project.org/, http://vegan.r-forge.r-project.org/

Peres-Neto PR, Olden JD, Jackson DA (2001) Environmentally constrained null models: site suitability as occupancy criterion. Oikos 93:110

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

Roberts DW, Wight D (1988) Plant community distribution and dynamics in Bryce Canyon National Park. United States Department of Interior National Park Service

Roberts DW (2006) LABDSV: Laboratory for Dynamic Synthetic Vegephenomenology. R package version 1.2–2. Available at http://cran.r-project.org/

Schnittler M, Unterseher M, Tesmer J (2006) Species richness and ecological characterization of myxomycetes and myxomycete-like organisms in the canopy of a temperate deciduous forest. Mycologia 98:223

Sidak Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62:626–633

Swan JMA (1970) An examination of some ordination problems by use of simulated vegetational data. Ecology 51:89–102

Whitehouse HE, Bayley SE (2005) Vegetation patterns and biodiversity of peatland plant communities surrounding mid-boreal wetland ponds in Alberta, Canada. Can J Bot 83:621–637

Electronic supplementary material

S1. Expected value of Beals smoothing for a “random” species

We demonstrate here that the expected value of b_{ij} for a “random” target species j in site i is the species relative frequency. Let $p_{j/k}$ be the (true) probability of species j conditioned to the appearance of species k and let the number of appearances of species n_k be a fixed quantity. Then, the number of observed joint occurrences between the two species, $m_{j/k}$, is a random variable (RV) distributed following a Binomial law, $m_{j/k} \sim Bin(n_k, p_{j/k})$, with mean $E(m_{j/k}) = n_k \cdot p_{j/k}$. Following this, $\hat{p}_{j/k}$ (the probability of species j conditioned to k estimated from n_k occurrences of species k) will be a RV with mean $E(\hat{p}_{j/k}) = p_{j/k}$. Now, let s_i be the number of species at a given location i (excluding the species of interest if present). s_i is also considered a fixed quantity. As b_{ij} is simply an average of $\hat{p}_{j/k}$ values, it is easy to obtain the mean of the RV b_{ij}:

$$E(b_{ij}) = E(\frac{1}{s_i}\sum_{k=1}^{s_i}\hat{p}_{j/k}) = \frac{1}{s_i}\sum_{k=1}^{s_i}E(\hat{p}_{j/k}) = \frac{1}{s_i}\sum_{k=1}^{s_i}p_{j/k}$$

However, if the target species j is a “random” species, meaning that it is completely unrelated to the reference species, the true conditioned probabilities are equal to the target species frequency (p_j). That is, $p_{j/k} = p_j$ for all species k. This straightforwardly yields $E(b_{ij}) = p_j$, as we wanted to demonstrate.

S2. Extending the Beals smoothing function to species abundance values

Although Beals smoothing was originally intended to be a transformation for binary species data tables, nothing prevents us from computing it using the information contained in the abundances of table **X**. Of course, this is done at the cost of making additional ecological assumptions. Since there are two vectors of parameters in eq. (2), such a generalization can be done in two corresponding ways, which are independent and compatible.

Perhaps the most natural generalization is to replace, in eq. (2), the vector of presence/absence values in the target sampling unit, $\mathbf{x}_i^0$, by abundance values, $\mathbf{x}_i$. With this substitution, the Beals smoothing function becomes a weighted average of estimated conditional probabilities, where the weights are the species abundances, and the resulting values become considerably smoother. This generalization implies the following ecological assumption: *The abundance values in a given sampling unit are related to the relative performance of the species under the environmental conditions of the sampled habitat.*

The second generalization consists in using abundance data for the computation of $\hat{\mathbf{p}}_{j/}$. Specifically, abundances of reference species *k* can be included as weights to assess the number of joint occurrences between species *k* and *j* (the target species): $m_{j/k} = \sum_{i=1}^{n} y_{ik} \cdot y_{ij}^0$. The vector of estimated conditional probabilities is then $\hat{\mathbf{p}}_{j/} = Diag(\mathbf{Y}^t\mathbf{Y}^0)^{-1}\mathbf{Y}^t\mathbf{y}_j^0$ and the interpretation of $\hat{p}_{j/k}$ is slightly different. If the abundance values are individual counts, then $\hat{p}_{j/k}$ is the estimated probability of “finding species *j* where an individual of species *k* has been found”. Generally speaking, the effect of including abundances in this way provides a “refined assessment” of the estimated conditional probability. It can be done assuming a different hypothesis for each reference species *k*: *The abundance values of species k (and not only its presence) can be predicted from the environmental conditions of the corresponding sampled sites*.

The abundance values of the target species do not play any role in any of these two generalizations if eq. (2) is used. As stated above, these two generalizations of Beals smoothing can be applied independently or simultaneously. That is, one could choose to keep the initial binary definition for $\hat{p}_{j/k}$ and use a weighted average for b_j; or instead use abundances for $\hat{p}_{j/k}$ while keeping the average unweighted; or else use abundances in both cases (i.e. using both generalizations).

Once a target species has been proven to be related to the main ecological patterns, another interesting ecological question is whether its abundance values can be modeled. The following simple test can be devised to address this question:

Beals species abundance (BSA) test:

- H_0: Abundances values **are not** related to the “sociological favorability” of the species.
- H_1: Abundances values **are** related to the “sociological favorability” of the species.

Answering this question affirmatively for species *k* would allow us to use its abundances when computing $\hat{p}_{j/k}$ for any other species *j*. A correlation measure appears naturally as a suitable test statistic. Such correlation analysis has to be restricted to those sampling units where the species has been found in order to avoid the zero-truncation problem. In addition, the permutation method has to be restricted to within those sampling units where the species has been found. As Beals smoothing function is independent of the target species abundance values, this restricted permutation method does not affect its value. Thus, the BSA test turns out to be a simple correlation test whose reference distribution is generated by permutations on one of the vectors. Naturally, if the number of species occurrences is very low – say less than 5 – this test will have very low statistical power.