Ecological Indicators 45 (2014) 721–727

Contents lists available at ScienceDirect

Ecological Indicators

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

Pairwise measures of species co-occurrence for choosing indicator species and quantifying overlap

Thomas M. Neeson[a,*], Yael Mandelik[b]

[a] Center for Limnology, University of Wisconsin, Madison, WI 53704, USA
[b] Department of Entomology, The Hebrew University of Jerusalem, PO Box 12, Rehovot IL-76100, Israel

ARTICLE INFO

Article history:
Received 4 March 2014
Received in revised form 7 May 2014
Accepted 3 June 2014

Keywords:
Bivariate covariance
Co-distribution
Mean pairwise index
Proportional similarity
Surrogate
Overdispersion
Community ecology

ABSTRACT

One of the most important ecological relationships between any two species is the degree of overlap in their distributions, i.e., their co-occurrence. Quantifying this relationship is a key step in the selection of indicator species and many other analyses in conservation biology and ecology. We derived a measure of the co-occurrence of two species based on the relative mutual information (RMI) of their distributions, and then compared its performance to three existing statistics: bivariate or binary covariance (BC), mean pairwise index (MPI), and proportional similarity (PS). To make this comparison, we measured co-occurrence values for all pairwise combinations of species collected from three communities (ground-dwelling beetles, moths, and vascular plants) in the Jerusalem Mountains and Judean Foothills, central Israel. We then used these co-occurrence values to address two different ecological problems: the challenge of identifying good indicator species, and the question of whether congeneric species co-occur more than species from different genera. We found that PS and RMI were the most reliable basis for choosing indicator species, but these two statistics differed in their error structures: PS had lower rates of type I errors (false positives), while RMI had lower rates of type II errors (false negatives). We also found that congeneric species co-occurred more often than species from different genera, but this pattern was statistically significant for only some of the measures of co-occurrence. In our analysis, then, the conclusion that we reached regarding the co-occurrence of congeneric species depended on which co-occurrence statistic was used. We therefore caution that available co-occurrence statistics should not be used interchangeably, because the ecological inferences drawn from a study may depend on the choice of co-occurrence statistic. In summary, we recommend PS and RMI as pairwise measures of species co-occurrence for investigating the reliability of biodiversity indicators and other applications in conservation biology and ecology.

1. Introduction

The measurement and analysis of species co-occurrence patterns is one of the most fundamental topics in ecology. In community ecology, co-occurrence patterns are often described at the community level, e.g., by summarizing the number of checkerboard units or other patterns in the community matrix (see Gotelli, 2000 for a review). In this paper, we focus on a different but related aspect of co-occurrence: the measurement of the degree of overlap in the distributions of two species (Veech, 2013, 2014). This problem entails quantifying the similarity of two rows in a standard community matrix, where each row represents a frequency distribution describing the presence or abundance of a species across sites. Though pairwise measures of species co-occurrence are less commonly used than community-level statistics, pairwise measurements are nonetheless an essential component of many analyses in community ecology (e.g., testing whether related species co-occur; Webb et al., 2002; Bell, 2005), conservation biology (e.g., the selection of indicator and surrogate species; Caro and O'Doherty, 1999; Caro, 2010), and biogeography (e.g., the identification of chorotypes or other bio-geographic units; Olivero et al., 2011).

Though several pairwise measures of co-occurrence exist, ecologists have little basis for choosing among these measures for the selection of indicator species and other related applications. Furthermore, given that the measurement of co-occurrence is akin to measuring the overlap of two frequency distributions, we hypothesized that measures of distributional similarity developed in other disciplines might also be useful for quantifying pairwise species co-occurrence. We were particularly interested in evaluating the

* Corresponding author. Tel.: +1 608 338 2235; fax: +1 608 265 2340.
E-mail address: neeson@wisc.edu (T.M. Neeson).

http://dx.doi.org/10.1016/j.ecolind.2014.06.006
1470-160X/

utility of mutual information, which is a mathematical measure of the amount of information that one statistical distribution provides about another (Manning and Schutze, 1999; Hart et al., 2000; MacKay, 2005). Given the widespread use of mutual information in other domains and its appealing mathematical properties (discussed later), we hypothesized that the mutual information of two species' distributions might be a useful measure of their co-occurrence.

In this paper, our goals were to (1) derive a measure of the co-occurrence of two species based on the relative mutual information (RMI) of their distributions, and (2) compare this statistic to three commonly used co-occurrence statistics: proportional similarity (PS; Schoener, 1970), mean pairwise index (MPI; Winston, 1995), and bivariate species covariance (BC; Bell, 2005). To make this comparison, we used these four measures of co-occurrence (RMI, BC, MPI, and PS) to address two different ecological problems: the challenge of identifying good indicator species, and the question of whether congeneric species tend to co-occur more often than species from different genera. We briefly introduce these two applications and our motivation for including them in this study.

The first application concerns the use of co-occurrence statistics in conservation biology for the selection of indicator or surrogate species. Conservation biologists use surrogate species in a variety of roles (Caro, 2010); we focus on the simplest type of indicative assessment, i.e., the use of the presence of one species as an indicator of the presence of a second, unobserved species. In this application, useful measures of co-occurrence are those that can be used to identify a species whose distribution is tightly coupled with that of another species. The majority of research in this field is empirical rather than theoretical, with highly variable results depending on the study system and scales (Hess et al., 2006; Larsen et al., 2009; Sartersdal and Gjerde, 2011). A generic measure of co-occurrence (i.e., not one derived from a specific data set) may thus afford a better basis for choosing and evaluating indicator and surrogate species.

The second application concerns the use of co-occurrence statistics to describe patterns in the co-occurrence of related species. Ecologists have long hypothesized that phylogenetic and taxonomic relationships among species can affect the structure of present-day communities (Johnson and Stinchcombe, 2007). These patterns come about because, in many cases, closely related species exhibit similar traits and habitat preferences (Johnson and Stinchcombe, 2007). If environmental filtering plays a dominant role in structuring communities, then the ecological similarity of related species might lead to a high degree of co-occurrence among these species (e.g., as in Webb, 2000; Tofts and Silvertown, 2000; Weiblen et al., 2006). On the other hand, if competition plays a more dominant role in structuring communities, then the ecological similarity of taxonomically-related species might lead to low co-occurrence among these species via competitive exclusion (e.g., as in Silvertown et al., 2001; Cavendar-Barres et al., 2006; Webb et al., 2006). Measures of co-occurrence clearly have a central role in this research area, and useful measures of co-occurrence hold the promise of providing key insights into the processes governing community assembly. Because existing measures of co-occurrence differ in their mathematical properties, we hypothesized that the choice of co-occurrence statistic might determine whether statistically significant patterns of clustering or overdispersion of related species are observed.

In this paper, we develop a measure of the co-occurrence of two species based on the relative mutual information of their distributions. To compare the usefulness of our new statistic to other, existing measures of species co-occurrence, we calculate the pairwise co-occurrence values for three different community datasets (beetles, moths, and plants, all from central Israel). We then use these co-occurrence values to investigate the two ecological questions introduced earlier: (1) Which measure of co-occurrence is the most reliable basis for choosing indicator species?, and (2) Do taxonomically-related species co-occur more often than unrelated species? Our motivation for the first analysis was the fact that indicator reliability could serve as an independent measure of how well each statistic described species co-occurrence. In the second analysis, our aim was to explore whether the ecological conclusion that we reached (i.e., whether related species co-occur) might depend on which co-occurrence statistic was used. If true, this analysis would serve as a cautionary reminder that co-occurrence statistics should not be used interchangeably in ecological analyses, and demonstrate which statistics show overlapping or dispersed results.

2. Methods

We first review existing pairwise measures of co-occurrence, and then derive a measure of co-occurrence based on the relative mutual information of two species' distributions. We then describe field survey data of beetles, moths and plants in the Jerusalem Mountains and Judean foothills, Israel, and our two research questions.

The co-occurrence of two species i and h is the similarity or overlap of the two vectors $(N_{i1},N_{i2},\ldots,N_{ir})$ and $(N_{h1},N_{h2},\ldots,N_{hr})$ in a standard community matrix, or a statistic derived from these two vectors. For presence–absence data, each element N_{ij} has value 1 if species i is present at site j, and value 0 if the species is absent. For abundance data, N_{ij} gives the number of individuals of species i at site j. Measures of co-occurrence based on the proportional distribution of each species can be calculated by first finding the total abundance of each species (i.e., the row totals) as

$$Y_i = \sum_{j=1}^{r} N_{ij}$$

and then the proportion of species i that occurs at location j as

$$p_{ij} = \frac{N_{ij}}{Y_i}$$

2.1. Existing measures of co-occurrence

We considered three existing measures of co-occurrence: proportional similarity (PS) (Schoener, 1970), bivariate species covariance (BC) (Bell, 2005), and mean pairwise index (MPI) (Winston, 1995). The proportional similarity (PS) of two species i and h over r sites is given by

$$PS_{ih} = 1 - 0.5\sum_{j=1}^{r} |p_{ij} - p_{hj}|$$

where p_{ij} is the proportion of species i at site j, and p_{hj} is the proportion of species h at site j.

The second statistic, binary or bivariate covariance (BC), is given by

$$BC\left(X_i, X_j\right) = \frac{(n_{11}n_{00} - n_{10}n_{01})}{r(r-1)}$$

where n_{11} is the number of sites (out of r total) with both species, n_{00} is the number of sites with neither species, and n_{10} and n_{01} are the number of sites with one species but not the other. Bell (2005) notes that BC is mathematically equivalent to the correlation coefficient given by Kershaw (1960) and to a modified version of the C score introduced by Stone and Roberts (1990).

The third statistic, mean pairwise index (MPI), is given by

$$\mathrm{MPI}_{ij} = \frac{n_{11}}{\min(n_{01}, n_{10})}$$

where n_{11} is the number of sites with both species, and n_{01} and n_{10} are the sites with only species i and j, respectively.

Three related measures, which we do not explore here, are the Sorenson–Dice index (Dice, 1945), the Bray–Curtis dissimilarity (Bray and Curtis, 1957) and the Jaccard index (Real and Vargas, 1996). Though these indices are mathematically similar to the indices examined here, they are typically used to measure the similarity of two sites' species compositions (Faith et al., 1987) rather than the co-occurrence of two species across sites.

2.2. A new measure of co-occurrence

We derived an index of co-occurrence based on the relative mutual information (RMI) of the two vectors describing the proportional distributions of species i and h, i.e., the RMI of the two vectors $S_i = (p_{i1}, p_{i2}, \ldots, p_{ir})$ and $S_h = (p_{h1}, p_{h2}, \ldots, p_{hr})$. We based our derivation on Colwell and Futuyma (1971), who used mutual information as the basis for an index of niche overlap. Here, we adapt their derivation to the problem of measuring the overlap of species across sites. Though our derivation is similar, it is simpler, because we do not need to account for the degree of distinctness of resource states.

To calculate the relative mutual information for species i and h, we first define the uncertainty (or entropy) in the spatial distribution of an individual of species i, which is

$$H(X) = -\sum_{j=1}^{r} p_{ij} \log p_{ij}$$

Similarly, the uncertainty for an individual of species h is

$$H(Y) = -\sum_{j=1}^{r} p_{hj} \log p_{hj}$$

where the logarithm can be taken to any base. Uncertainty has an intuitive meaning in these equations. For species i, $H(X)$ has a maximum value when the proportional distribution of species i is uniform across all locations; in other words, when there is maximum uncertainty as to the location of a randomly-selected individual of species i. $H(X)$ has a minimum value of 0 when species i occurs at only one location; in this case, there is no uncertainty as to the location of a randomly-selected individual of species i.

Next we define a new vector by letting $t_j = p_{ij} + p_{hj}$. The vector $(t_1, t_2, \ldots, t_r)$ describes the proportional distribution of individuals of both species across all r locations. The total uncertainty of any one individual is then

$$H(XY) = -\sum_{j=1}^{r} t_j \log t_j$$

If the proportional distribution of species i is known, then we can subtract the uncertainty associated with species i, $H(X)$, from the total uncertainty, $H(XY)$. The remainder is the amount of uncertainty of an individual of species h with respect to location, given that the distribution of species i is known:

$$H(Y|X) = H(XY) - H(X)$$

A measure of how much information the distribution of species i provides about the distribution of species h is the mutual information, which is the uncertainty in Y minus the uncertainty in Y given that X is known:

$$I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$$

The limits of $I(X;Y)$ are 0 (when i and h never co-occur) and, for presence-absence data, $2\log_2(2)$, when the proportional distributions of species i and h are identical.

A more useful measure of co-occurrence is the relative or normalized mutual information, which results from rescaling $I(X;Y)$ to have an upper bound of 1:

$$R(X;Y) = \frac{1}{2\log_2(2)}\,[H(X) + H(Y) - H(XY)]$$

For abundance data, the scaling constant $1/2\log_2(2)$ should be replaced by the term $1/[H(X)+H(Y)]$.

2.3. Mathematical properties of co-occurrence statistics

The four co-occurrence measures under consideration here differ in their mathematical properties. Only PS and RMI have a standardized range of (0,1), can accommodate abundance data, and do not depend per se on the commonness or rarity of the two species (Colwell and Futuyma, 1971; Feinsiner et al., 1981). BC, PS, and RMI are all symmetric (e.g., $\mathrm{PS}_{i,h} = \mathrm{PS}_{h,i}$), but MPI is not.

2.4. Data: Beetles, moths and plants

We used field collections of ground-dwelling beetles, moths and vascular plants obtained from 40 1000 m^2 plots in the Jerusalem Mountains and the Judean Foothills, a dry Mediterranean ecosystem, central Israel. Plots were located in the main natural habitats found in the region—dwarf shrubland, Mediterranean shrubland, open and dense chaparral, and represented main environmental variation in the region (details in Mandelik et al., 2012).

Beetles were collected with pitfall traps at all 40 sites, 12 traps per plot. Each plot was sampled five times during 2001–2002, in the winter, early and late spring, summer and autumn; traps were left open for one week during each sampling round. Moths were collected at 25 sites using light traps, one trap per plot. Each plot was sampled six times during 2001–2002, in the winter, early and late spring, summer and early and late autumn; traps were opened for one night during each sampling round, from dusk to dawn. Vascular plants in each plot were recorded along four 50 m transects twice during the spring of 2002. Sampling effort required for the faunal and floral surveys was established in a preliminary study using species accumulation curves (details in Mandelik et al., 2007).

These three communities differed in species richness, rarity and abundance, enabling us to explore the possible effects of these characteristics on the performance of the co-occurrence measures. Species richness among beetles ($n = 424$ species) and plants (420 species) was nearly four times as high as species richness among moths (111 species; Fig. 1). The moth community exhibited many common species, while the beetle community was characterized by many rare species. The plant community was nearly as species-rich as the beetle community but contained many more common species and fewer rare species.

2.5. Question 1: Which measure of co-occurrence is the most reliable basis for choosing indicator species?

At any one location, the relationship between an indicator and its target species will be one of four types: true positive (locations where both the indicator and target species are present; TP), true negative (neither species is present; TN), false positives (the indicator species is present, but the target is not; FP), and false negatives (the indicator species is absent, but the target is present; FN). The percentage of locations that are true positives or true negatives gives the percentage of correct indications. The percentage of false positives is an estimate of the type I error rate,

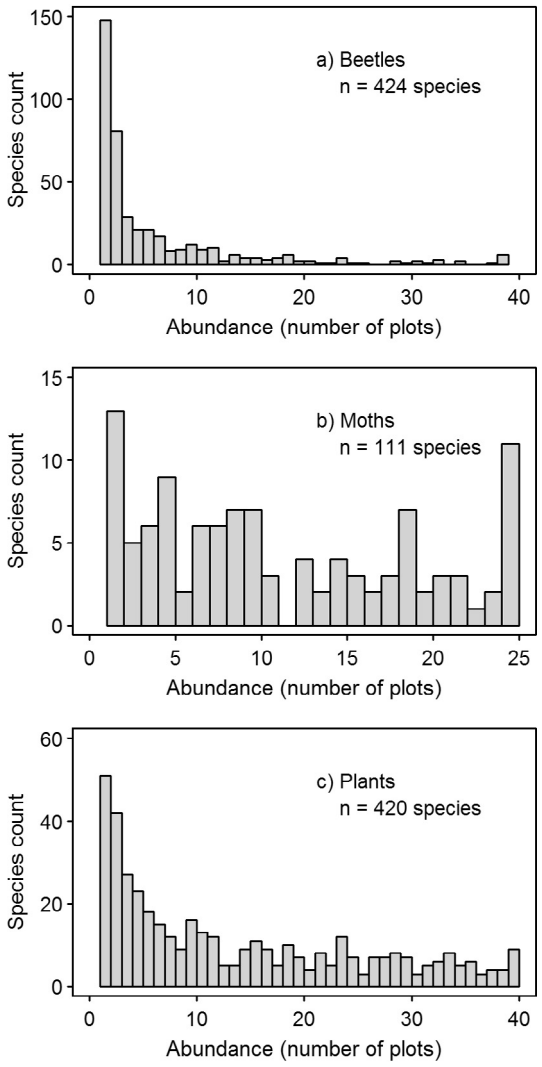

Fig. 1. Species abundance distributions for collections of (a) beetles, (b) moths, and (c) plants from the Jerusalem Mountains and Judean Foothills, Israel.

while the percentage of false negatives is an estimate of the type II error rate. These two error rates are of practical interest for the selection of indicator species and other applications. In some instances it may be desirable to choose an indicator with a small type I error rate (e.g., in order to minimize false positives when selecting an indicator of a rare or endangered species). In other instances it may be desirable to choose an indicator with a small type II error rate (e.g., in order to minimize false negatives in the detection of an invasive or exotic species).

We compared the ability of all four measures of species co-occurrence to serve as a method for selecting indicator species by comparing their average performance within each of the three datasets (beetles, moths, and plants). To do this, we performed the following analysis for each measure of co-occurrence, and for each of the three datasets. For each species in a dataset, we calculated its co-occurrence value with all other species in that dataset. For each species, we then selected another species to serve as an indicator of that species by choosing the species with the highest co-occurrence value. For this indicator and target species pair, we then tallied the number of TP, TN, FP, and FN across all sites. To summarize the average performance of each co-occurrence statistic, we then calculated the mean number of TP, FN, FP, and FN in each dataset.

In this analysis, our aim was to use TP, TN, FP, and FN as a set of metrics to describe the properties (i.e., accuracy and type I and type II error rates) of each of the four co-occurrence measures. We reasoned that a meaningful measure of co-occurrence is one which reports high levels of co-occurrence for two species with a large number of TP and TN, and which reports low levels of co-occurrence for two species with few TP and TN. Conversely, a measure of co-occurrence is less meaningful if it reports high values of co-occurrence for two species with few TP and TN. In the analysis performed here, the species chosen as indicators are those with the highest measured co-occurrence according to each of the four measures of co-occurrence (PS, BC, MPI, RMI). By reporting average TP, TN, FP, and FN for each measure of co-occurrence, we can comment on whether a particular measure of co-occurrence is meaningful or not.

2.6. Question 2: Do related species co-occur more often than unrelated species?

We tested (a) whether congeneric species tended to co-occur more often than species from different genera, and (b) whether the statistical significance of this pattern differed among the four co-occurrence statistics under consideration. To do this, we performed the following analysis for each measure of co-occurrence, and for each of the three datasets. First, we measured co-occurrence for each of the pairwise combinations of species in the dataset. We then grouped together all cases in which both species belonged to the same genus, and grouped separately all cases in which the two species did not belong to the same genus. We then performed a *t*-test to determine whether the co-occurrence values among species from the same genus were larger than those among unrelated species. The null hypothesis in these comparisons is that taxonomic relatedness has no effect on likelihood of co-occurrence. We used taxonomic identity as a proxy for phylogenetic relatedness, though we recognize that the quantitative evolutionary distance between species may vary widely among genera or families.

We used the R statistical package (R Development Core Team, 2012) for all analyses.

3. Results

3.1. Question 1: Which measure of co-occurrence is the most reliable basis for choosing indicator species?

We found consistent differences in the ability of the four co-occurrence measures to serve as a basis for selecting indicator species. The four measures of co-occurrence were ranked by accuracy (i.e., TP + TN/number of sites) in the same order for all three datasets: PS and RMI were the most accurate, followed by BC and then MPI (Table 1). The relative differences in their performance were also similar among datasets. RMI performed nearly identically to PS, and the difference was never statistically significant (*t*-test, $P > 0.05$ for beetles, moths and plants). Their overall accuracy (TP + TN/number of sites) ranged between ca. 87% and 92%. The differences between BC and PS, BC and RMI, MPI and PS, and MPI and RMI were larger and statistically significant (*t*-test, $P < 0.05$ for all three data sets). BC's accuracy ranged between ca. 81% and ca. 88%, while MPI performed relatively poorly (ca. 58–72% accuracy).

Type I and type II error rates also differed among the four measures (Table 1). Although PS and RMI had similar accuracy overall, RMI always had a higher type I error rate, while PS had a higher type II error rate. MPI had the most unusual error structure, notably a very high type I error rate (greater than 25% for all datasets).

Scatterplots of individual pairwise co-occurrence measurements show broad similarity among some measures of co-occurrence, and broad differences among others (Fig. 2). PS and RMI behaved similarly and were strongly correlated ($r = 0.96$ for

Table 1
Mean number of sites in which the indicator and target species are both present (true positives, TP), neither species is present (true negative, TN), only the indicator is present (false positives, FP), and only the target is present (false negatives, FN). For each dataset, measures of co-occurrence are ranked according to percent correct (PC), which is TP + TN divided by the total number of sites.

	TP	TN	FP	FN	PC (%)
Beetles					
Proportional similarity (PS)	4.24	32.56	1.70	1.51	92.00
Relative mutual information (RMI)	4.73	31.88	2.28	1.01	91.53
Bivariate covariance (BC)	4.23	30.85	3.40	1.51	87.70
Mean pairwise index (MPI)	4.76	24.07	10.18	0.98	72.09
Moths					
Proportional similarity (PS)	9.66	12.28	1.34	1.41	91.41
Relative mutual information (RMI)	10.18	11.75	2.44	0.63	91.38
Bivariate covariance (BC)	7.43	12.19	2.00	3.38	81.76
Mean pairwise index (MPI)	8.05	5.78	8.41	2.77	57.62
Plants					
Proportional similarity (PS)	10.80	23.93	2.90	2.37	86.83
Relative mutual information (RMI)	11.80	22.82	4.01	1.37	86.55
Bivariate covariance (BC)	9.91	22.61	4.22	3.25	81.31
Mean pairwise index (MPI)	11.72	11.80	15.03	1.44	58.80

beetles, $r = 0.97$ for moths and plants). BC exhibited moderate correlation with RMI ($r = 0.68$ for beetles, $r = 0.50$ for moths, $r = 0.47$ for plants) and with PS ($r = 0.63$ for beetles, $r = 0.45$ for moths, $r = 0.42$ for plants). MPI exhibited the lowest similarity with the other three measures, with correlation coefficients ranging from 0.11 to 0.56.

3.2. Question 2: Do related species co-occur more often than unrelated species?

We found that taxonomically-related species (i.e., congenerics) tended to co-occur more often than unrelated species, but the statistical significance of this pattern varied among datasets and among measures of co-occurrence (Table 2). The pattern was strongest for beetles, where the measured co-occurrence of related species was significantly higher than that of unrelated species for all four measures of co-occurrence. In plants, there was no statistically significant difference in the degree of co-occurrence of related and unrelated species. In moths, however, the statistical significance of the co-occurrence of related species depended on which co-occurrence measure was used. Species from the same genus co-occurred significantly more often when co-occurrence was measured by BC and RMI, but this pattern was not significant when co-occurrence was measured by MPI and PS.

Table 2
Mean co-occurrence values between related species (i.e., pairs of species from the same genus) and unrelated species and *P*-values for a *t*-test for whether related species co-occur significantly more often than unrelated species. Bold values indicate significant results for $P \leq 0.05$.

	Related species	Unrelated species	*P*
Beetles			
Proportional similarity (PS)	0.134	0.0649	$\leq \mathbf{0.05 \times 10^{-10}}$
Relative mutual information (RMI)	0.185	0.0998	$\leq \mathbf{0.05 \times 10^{-10}}$
Bivariate covariance (BC)	8.89×10^{-3}	1.33×10^{-3}	$\leq \mathbf{0.05 \times 10^{-9}}$
Mean pairwise index (MPI)	0.890	0.299	$\leq \mathbf{0.05 \times 10^{-3}}$
Moths			
Proportional similarity (PS)	0.313	0.269	>0.05
Relative mutual information (RMI)	0.419	0.366	$\leq$**0.05**
Bivariate covariance (BC)	1.15×10^{-3}	3.70×10^{-3}	$\leq$**0.05**
Mean pairwise index (MPI)	1.80	1.22	>0.05
Plants			
Proportional similarity (PS)	0.181	0.173	>0.05
Relative mutual information (RMI)	0.245	0.245	>0.05
Bivariate covariance (BC)	1.81×10^{-3}	1.70×10^{-3}	>0.05
Mean pairwise index (MPI)	1.47	1.39	>0.05

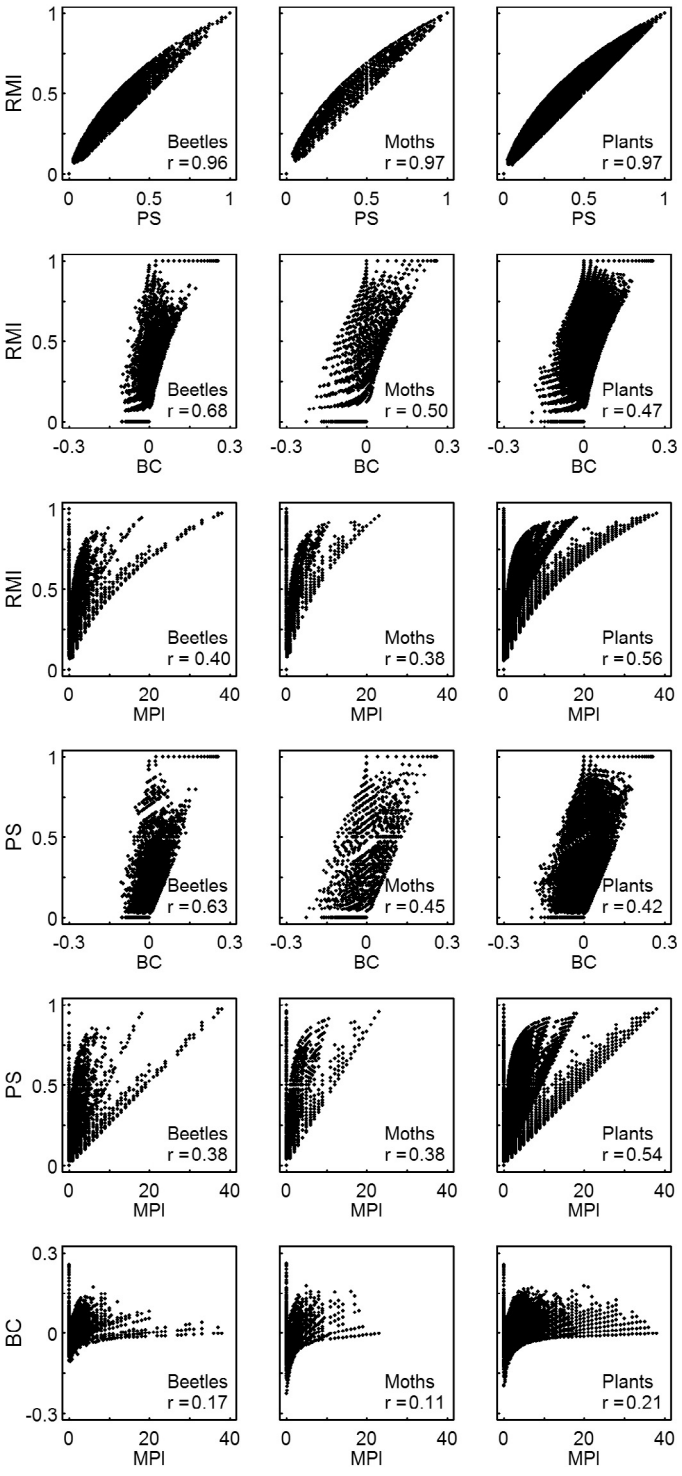

Fig. 2. Scatterplots and measured correlations (inset) between individual measurements of pairwise (species-species) co-occurrence for all combinations of the four different measures of co-occurrence (PS, RMI, BC, and MPI), for beetles (left column), moths (center column) and plants (right column).

4. Discussion

We found a clear basis for recommending PS and RMI as the two most useful pairwise measures of species co-occurrence. In our first analysis (Question 1), we used indicator reliability as an independent measure of how well each statistic described species co-occurrence. PS and RMI consistently outperformed BC and MPI across each of the data sets in our analysis, suggesting that PS and

RMI capture the most information about species' distributions and are the two most meaningful measures of co-occurrence. We recommend that MPI in particular should be avoided as a measure of co-occurrence. This statistic performed poorly on all three data sets in our analysis, and exhibited an extremely high type I error rate (i.e., indicators chosen using MPI resulted in many false positives).

Though PS and RMI were highly correlated, we found that even slight differences in the measured value of co-occurrence could determine the ecological inferences that we drew. As a result, in our second analysis (Question 2), we could not always reach a consistent conclusion regarding the co-occurrence of congeneric species. When we measured the co-occurrence of moths using RMI, we found a statistically significant difference between the degree of co-occurrence of congeneric and non-congeneric species. When we measured the co-occurrence of moths using PS, however, this pattern was not statistically significant. We do not have any basis for assuming that one of these conclusions about the co-occurrence of moths is more correct than the other, because there is no a priori reason to suspect that PS or RMI is the more ecologically meaningful measure of co-occurrence. Rather, RMI and PS simply describe different (though related) aspects of the overlap of species distributions. Because of these differences, PS and RMI should be considered meaningful but complimentary measures of the overlap of species distributions.

PS and RMI are both generic measures of species overlap that should be broadly applicable in biogeography (e.g., the identification of chorotypes; Olivero et al., 2011) and in conservation biology to the selection of indicators, surrogates, and umbrella species. The majority of research on the selection of indicator and surrogate species has been empirically derived, leading to variable and conflicting results on surrogate reliability, depending on the study system and scales (reviewed in Caro and O'Doherty, 1999; Caro, 2010). Generic measures of co-occurrence such as PS and RMI, on the other hand, provide a standardized metric for describing the degree of overlap of two species' distributions and so a way to quantify the performance of an indicator. Importantly, both measures have a standardized range of (0,1) and do not depend, per se, on the overall abundance of the species nor the number of sites, as the other measures tested.

Indicator species selected using PS and RMI were very accurate overall, ranging from 86% correct (for moths) to 92% correct (for beetles). In our data sets, then, pairwise measures of species co-occurrence were a reliable and useful basis for selecting indicator species. The differing type I and type II error structures of PS and RMI may also be of practical value. When choosing an indicator for a rare or threatened species, for example, it would be preferable to use a co-occurrence measure that is both reliable and has a low type I (false positive) error rate. For this application, PS may be most appropriate. In other instances it may be preferable to choose an indicator using a measure that is reliable and gives low type II (false negative) error rates, e.g., for invasive or exotic species. For this application, RMI may be most appropriate.

Though our primary goal was to develop and evaluate measures of co-occurrence for the selection of indicator species, our analysis also revealed interesting variation among taxa in the degree to which taxonomically-related species co-occurred. We found strong evidence that taxonomically-related species of beetles co-occurred, some evidence that related moths co-occurred, and no significant evidence that related plants co-occurred. One hypothesis for these differences is that more mobile species, such as beetles and moths, may have greater freedom to escape localized competitive effects, thus enabling co-existence with ecologically similar related species. Alternatively, communities of beetles, moths and plants may simply differ in the relative intensity or spatial scales of habitat filtering and competition. Disentangling these hypotheses and ascribing co-occurrence patterns to one particular mechanism is notoriously challenging (Losos, 2008). In plants, for example, there is some evidence that related species exhibit a high degree of co-occurrence (Tofts and Silvertown, 2000; Webb, 2000), as well as evidence that they do not (Silvertown et al., 2001; Cavendar-Barres et al., 2006; Webb et al., 2006). Webb et al. (2006) suggest that these patterns can in fact differ with spatial scale and life history stage.

Regardless of the underlying mechanisms or drivers of co-occurrence, information on the co-occurrence of taxonomically-related species is of practical value for biologists interested in using higher taxon surrogates, i.e., the use of families or genera as surrogates for species (Williams and Gaston, 1994; Balmford et al., 1996). These surrogates are widely used in ecology and paleontology (Caro, 2010), but their reliability is hypothesized to be influenced by the degree to which species from the same taxa co-occur (Balmford et al., 1996; Neeson et al., 2013; van Rijn et al., in press). Interestingly, our results provide some anecdotal support for this hypothesis. Mandelik et al. (2007), using our same dataset, found that higher taxon surrogates were most reliable for beetles, the taxa we found to have the highest co-occurrence of related species. Mandelik et al. (2007) found that higher taxon surrogates performed less well for plants and moths, taxa where we found less evidence for the co-occurrence of related species. More thorough tests of this hypothesis, using PS and RMI, should be a straightforward extension of studies on the reliability of higher taxon surrogates.

5. Conclusions

We recommend PS and RMI as pairwise measures of species co-occurrence. These two statistics exhibit different type I and type II error structures, but we cannot otherwise recommend one over the other. Indicator species chosen using these two measures were generally very accurate (>85% in all cases), suggesting that pairwise measures of species co-occurrence can be a reliable and useful basis for selecting indicator and surrogate species. PS and RMI are both generic measures of co-occurrence (i.e., not derived from a particular study system) and should therefore be broadly applicable across taxa and biogeographic settings to a diverse set of problems in conservation biology and ecology.

Acknowledgments

We thank M. Coll, I. Van Rijn, P. McIntyre, B. Peckarsky, the McIntyre lab at UW, and two anonymous reviewers for thoughtful comments on the ideas in this paper.

References

Balmford, A., Jayasuriya, A.H.M., Green, M.J.B., 1996. Using higher-taxon richness as a surrogate for species richness: II. Local applications. Proc. R. Soc. Lond., Ser. B, Biol. Sci. 263, 1571–1575.
Bell, G., 2005. The co-distribution of species in relation to the neutral theory of community ecology. Ecology 86, 1757–1770.
Bray, J.R., Curtis, J.T., 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27, 325–349.
Caro, T.M., O'Doherty, G., 1999. On the use of surrogate species in conservation biology. Conserv. Biol. 14, 1580–1591.
Caro, T.M., 2010. Conservation by Proxy. Island Press, Washington, DC.
Cavendar-Barres, J., Keen, A., Miles, B., 2006. Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. Ecology 87, S109–S122.
Colwell, R.K., Futuyma, D.J., 1971. On the measurement of niche breadth and overlap. Ecology 52, 567–576.
Dice, L.R., 1945. Measures of the amount of ecological association between species. Ecology 26, 297–302.
Gotelli, N.J., 2000. Null model analysis of species co-occurrence patterns. Ecology 81, 2606–2621.
Faith, D.P., Minchin, P.R., Belbin, L., 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69, 57–68.
Feinsiner, P., Spears, E.E., Poole, R.W., 1981. A simple measure of niche breadth. Ecology 62, 27–32.

Hart, P.E., Stork, D.G., Duda, R.O., 2000. Pattern Classification. Wiley-Interscience, New York, NY.
Hess, G.R., Bartel, R.A., Leidner, A.K., Rosenfeld, K.M., Rubino, M.J., Snider, S.B., Ricketts, T.H., 2006. Effectiveness of biodiversity indicators varies with extent, grain, and region. Biol. Conserv. 132, 448–457.
Johnson, M.T.J., Stinchcombe, J.R., 2007. An emerging synthesis between community ecology and evolutionary biology. Trends Ecol. Evol. 22, 250–257.
Kershaw, K.A., 1960. The detection of pattern and association. J. Ecol. 48, 233–242.
Larsen, F.W., Bladt, J., Rahbek, C., 2009. Indicator taxa revisited: useful for conservation planning? Divers. Distrib. 15, 70–79.
Losos, J.B., 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecol. Lett. 11, 995–1007.
MacKay, D.J.C., 2005. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge.
Manning, C.D., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge.
Mandelik, Y., Dayan, T., Chikatunov, V., Kravchenko, V., 2007. Reliability of a higher-taxon approach to richness, rarity and composition assessments at the local scale. Conserv. Biol. 21, 1506–1515.
Mandelik, Y., Dayan, T., Chikatunov, V., Kravchenko, V., 2012. The relative performance of taxonomic vs. environmental indicators for local biodiversity assessment: a comparative study. Ecol. Indic. 15, 171–180.
Neeson, T.M., Van Rijn, I., Mandelik, Y., 2013. How taxonomic diversity, community structure and sample size determine the reliability of higher taxon surrogates. Ecol. Appl. 23, 1216–1225.
Olivero, J., Real, R., Marquez, A.L., 2011. Fuzzy chorotypes as a conceptual tool to improve insight into biogeographic patterns. Syst. Biol. 60, 645–660.
R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ⟨http://www.R-project.org/⟩.
Real, R., Vargas, J.M., 1996. The probabilistic basis of Jaccard's index of similarity. Syst. Biol. 45, 380–385.
Sartersdal, M., Gjerde, I., 2011. Prioritising conservation areas using species surrogate measures: consistent with ecological theory? J. Appl. Ecol. 48, 1236–1240.
Schoener, T.W., 1970. Nonsynchronous spatial overlap of lizards in patchy habitats. Ecology 51, 408–418.
Silvertown, J., Dodd, M., Growing, D., 2001. Phylogeny and the niche structure of meadow plant communities. J. Ecol. 89, 428–435.
Stone, L., Roberts, A., 1990. The checkerboard score and species distributions. Oecologia 85, 74–79.
Tofts, R., Silvertown, J., 2000. A phylogenetic approach to community assembly from a local species pool. Proc. R. Soc. Lond., Ser. B, Biol. Sci. 267, 363–369.
van Rijn, I., Neeson, T.M., Mandelik, Y.M., 2014. Reliability and refinement of the higher taxa approach for bee richness and composition assessments. Ecol. Appl., http://dx.doi.org/10.1890/13-2380.1, in press.
Veech, J.A., 2013. A probabilistic model for analysing species co-occurrence. Global Ecol. Biogeogr. 22, 252–260.
Veech, J.A., 2014. The pairwise approach to analysing species co-occurrence. J. Biogeogr., http://dx.doi.org/10.1111/jbi.12318.
Webb, C.O., 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am. Nat. 156, 145–155.
Webb, C.O., Ackerly, D.D., McPeek, M.A., Donoghue, M.J., 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33, 475–505.
Webb, C.O., Gilbert, G.S., Donoghue, M.J., 2006. Phylodiversity-dependent seedling mortality, size structure, and disease in a Bornean rain forest. Ecology 87, s123–s131.
Weiblen, G.D., Webb, C.O., Novotny, V., Basset, Y., Miller, S.E., 2006. Phylogenetic dispersion of host use in tropical insect herbivore community. Ecology 87, s62–s75.
Williams, P.H., Gaston, K.G., 1994. Measuring more of biodiversity: can higher-taxon richness predict wholesale species richness? Biol. Conserv. 67, 211–217.
Winston, M.R., 1995. Co-occurrence of morphologically similar species of stream fishes. Am. Nat. 145, 527–545.