
Article:
Chetcuti, J orcid.org/0000-0002-1954-6105, Kunin, WE orcid.org/0000-0002-9812-2326 and Bullock, JM (2019) A weighting method to improve habitat association analysis: tested on British carabids. Ecography, 42 (8). pp. 1395-1404. ISSN 0906-7590

https://doi.org/10.1111/ecog.04295

A weighting method to improve habitat association analysis: tested on British carabids

Jordan Chetcuti[1,2], William E. Kunin[2] and James M. Bullock[1]

[1]NERC Centre for Ecology and Hydrology, Benson Lane, Wallingford Oxfordshire OX10 8BB, UK
[2]Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT UK

Corresponding author: Jordan Chetcuti, NERC Centre for Ecology and Hydrology, Benson Lane, Wallingford Oxfordshire OX10 8BB, UK. E-mail: jorche@ceh.ac.uk

Decision date: 13-Mar-2019

Abstract

Analysis of species' habitat associations is important for biodiversity conservation and spatial ecology. The original phi coefficient of association is a simple method that gives both positive and negative associations of individual species with habitats. The method originates in assessing the association of plant species with habitats, sampled by quadrats. Using this method for mobile animals creates problems as records often have imprecise locations, and would require either using only records related to a single habitat or arbitrarily choosing a single habitat to assign.

We propose and test a new weighted version of the index that retains more records, which improves association estimates and allows assessment of more species. It weights habitats that lie within the area covered by the species record with their certainty level, in our case study, the proportion of the grid cell covered by that habitat.

We used carabid beetle data from the National Biodiversity Network atlas and CEH Land Cover Map 2015 across Great Britain to compare the original method with the weighted version. We used presence-only data, assigning species absences using a threshold based on the number of other species found at a location, and conducted a sensitivity analysis of this threshold. Qualitative descriptions of habitat associations were used as independent validation data.

The weighted index allowed the analysis of 52 additional species (19% more) and gave results with as few as 50 records. For the species we could analyse using both indices, the weighted index explained 70% of the qualitative validation data compared to 68% for the original, indicating no accuracy loss.

The weighted phi coefficient of association provides an improved method for habitat analysis giving information on preferred and avoided habitats for mobile species that have limited records, and can be used in modelling and analysis that directs conservation policy and practice.
Key-words: carabids, Coleoptera, ground beetles, habitat classification, habitat preference, invertebrate, land cover, site fidelity, phi coefficient of association

Introduction

Habitat association analysis is used in determining the likely habitat requirements of individual species (Cole et al. 2010). These requirements are important, for example, in studying impacts of habitat loss and fragmentation (Maclean et al. 2011), dispersal and habitat connectivity (Brodie et al. 2016), and modelling foraging and movement over landscapes, such as in pollinator models (Lonsdorf et al. 2009) and conservation prioritisation (Pouzols and Moilanen 2014). Such analyses are particularly important when planning landscapes for conservation: for example, in assessing the impact of adding a patch of habitat for certain species, it is also necessary to understand which species avoid that habitat. Lawton et al. (2010) highlight that the approaches available for designing ecological networks are limited by the availability of evidence, usually using expert consensus. Habitat association analysis contributes to this evidence base.

Searching the literature for habitat association or preference returns many examples of species distribution models (SDMs) and indicator species analysis. Examples of analysis looking at preference of a species to each of several alternative habitats are returned less often. For example, SDMs predict where species are likely to be found within a landscape, with habitat type being only one factor (De Lima et al. 2016). Indicator species analysis identifies species that best represent a habitat or group of habitats, and is used in monitoring habitat condition (Hill et al. 1975, De Gasperis et al. 2016). Direct analysis of which habitats a species prefers and which it avoids, which is particularly useful in conservation planning, are few. In this paper, we consider a direct approach to determine habitat association, which comprises the relative preference of a species for multiple habitats.

Information on habitat associations is generally derived from expert knowledge (Lonsdorf et al. 2009) or analysis over a small geographic area (Ball et al. 2013, De Lima et al. 2016, Ferrão et al. 2018) and is often limited to associations with a single habitat or a few broad habitats (Webb et al. 2017). Large-scale analysis of habitat association de-emphasises the less frequent recordings of a species in a habitat in which the species is transient, which could be misconstrued at a small-scale as association. Although SDMs (Petit et al. 2003, Phillips et al. 2006, Porto et al. 2018) and indicator species analyses (Hill et al. 1975, Gardner 1991, Ricotta et al. 2015) are often done over large scales, this is rare for analysis of the preference of a species . Exceptions are Eyre and Luff (2004), who used ordination to study habitat preferences of carabids in North East England and the Scottish Border, and Redhead et al.(2016) who used general linear mixed effects models to study butterfly habitats across Britain.

Eyre and Luff (2004) used ordination in a straightforward way, giving each carabid species a weighted average from positive to negative for each habitat. They did, however, point out that care should be taken in interpreting their findings due to some anomalous results. Redhead et al. (2016) used the coefficients from their model to derive associations. Their method worked well, albeit with large variation in the associations within individual species, but needed approximately 5000 records to ensure accuracy. They used this approach, as other methods required more precise locations information than the 1 km they used.

De Cáceres and Legendre (2009) created a framework for ecologists explaining when to use IndVal or an alternative, the Phi coefficient of correlation (Pearson 1896). We focus in this paper on the Phi coefficient of correlation, ("correlation index") which like IndVal is simpler than ordination. Unlike IndVal, the correlation index gives a negative association value when a species appears to avoid a habitat, and uses species' absences to provide extra information (De Cáceres and Legendre 2009). The Phi coefficient gives degree of preference for a habitat compared to other groups. By contrast IndVal assesses how much the target site group matches a set of sites where the species is found and is an indicator species analysis. The correlation index was created by Karl Pearson (1896) and at its simplest is the binary version of the Pearson's correlation (De Cáceres et al. 2008). It is the preferred method in plant science for calculating site fidelity (De Cáceres and Legendre 2009), but has not been adopted more generally despite De Cáceres and Legendre's (2009) framework. The index uses two binary vectors to describe a location: one representing presence or absence of the species and the other whether a location is the habitat of interest. The index does not incorporate uncertainty in the habitat of the location, it is either habitat or not. Species records often have a degree of uncertainty, particularly concerning the spatial resolution of the record. The area covered by the resolution of the record may contain multiple habitats. The binary nature of the correlation index requires either removal of mixed or uncertain habitat data or a judgement as to which habitat to assign. While this might be considered as an error in the record, movement of individuals from preferred into adjacent less-preferred habitats is common (Ries et al. 2004), and so the precise location in which a mobile individual is found may not be in a preferred habitat. To incorporate these issues, we propose a new version of the correlation index, adding a third vector to each record, which is a weighting based on the certainty of the habitats at a location. We present the weighting as the proportion of a

particular habitat in 100 m grid cell. However, the weighting could be the probability of correctly classifying a habitat from remote sensing or a combination of weightings.
In this paper, we present our weighted version of the correlation index and test it against the original version using a case study of carabid beetles of Great Britain. We also carry out a partial validation of the correlation indices using qualitative data from species descriptions. The analysis uses records from the UK National Biodiversity Network (NBN) Atlas (2018) and Centre for Ecology and Hydrology (CEH) Land Cover Map 2015 (LCM2015) (Rowland et al. 2017a). We used a method that considers the number of other species within the family found at a location as proxy for survey effort (Hickling et al. 2006, Redhead et al. 2016). We use an absence threshold of 14 carabid species and conduct a sensitivity analysis of the threshold value. Most species have fewer than 1000 records. We, therefore, ascertain how many records are required to give a valid estimate of habitat association.

Methods

Correlation indices

The original correlation index uses binary presence-absence with each location assigned to one group (habitat) (De Cáceres and Legendre 2009). The index is the Pearson correlation coefficient for two binary vectors with length L, one vector representing the species presence/absence at each location ($\boldsymbol{s}$) and another representing if each location is the habitat of interest ($\boldsymbol{h}$) (De Cáceres and Legendre 2009). The lengths and sums of each vector are used in equation 1

$$\Phi = \frac{N \times n_p - n \times N_p}{\sqrt{(N \times n - n^2) \times \left(N \times N_p - N_p^2\right)}} \quad (1)$$

Where N is the total number of locations ($N = L$), N_p the number of locations with the habitat of interest ($N_p = \sum_{i=1}^{L} h_i$), n is total number of occurrences across all locations ($n = \sum_{i=1}^{L} s_i$), and n_p is the number of occurrences in habitat of interest ($n_p = \sum_{i=1}^{L} h_i s_i$). In the event that a location is not a point location and instead covers an area, a location could contain more than one habitat. For example, in location 4 (Table 1), an area location contains acid grassland (2%), inland rock (59%) and heather (39%). We do not know in which habitat the species was found, therefore when calculating the original index, either only locations that contain a single habitat could be included or a habitat would need to be chosen. We might choose to discard all locations with more than one habitat. This would leave locations 1, 2, 3, 5, and 7 in Table 1. The carabid species of interest is then either present or absent within that single land cover type. Using this approach can remove a large proportion of the data, sometimes making a species unviable for analysis. Another way of conducting the unweighted analysis would have been to choose the habitat covering the largest proportion of the 100 m location; a version of the analysis doing this can be found in Appendix 1. To allow the use of a larger proportion of the data, we created a weighted version of the index (equation 2).

$$\Phi^w = \frac{N \times n_p^w - n \times N_p^w}{\sqrt{(N \times n - n^2) \times \left(N \times N_p^w - N_p^{w2}\right)}} \quad (2)$$

This version added a third non-binary vector of the weighting of each habitat at each location ($\boldsymbol{w}$). This weighting could be any by which each location sums to one (for example land cover classification certainty) but we used the proportion of each habitat. All three vectors have length L. N is still the total number of locations($N = \sum_{i=1}^{L} w_i$)r, and n is still the total number of occurrences across all locations($n = \sum_{i=1}^{L} w_i s_i$). The values of N and n are the same as they would be if each of the locations only had a weighting of one (a single habitat in our example). The N_p and n_p values change however, now denoted as N_p^w and n_p^w. These can be calculated as $N_p^w = \sum_{i=1}^{L} w_i h_i$ (lower than a hypothetical N_p would be) and $n_p^w = \sum_{i=1}^{L} w_i h_i s_i$(smaller than a hypothetical n_p). So for only the data in Table 1 (assuming no threshold was applied) and with Inland rock being the habitat of interest; $n_p^w = 0.59 + 1 = 1.59$, $N_p^w = 0.59 + 1 + 1 = 2.59$, N = 7 and n = 2 and therefore using the equation for the weighted correlation index (equation 2) gives;

$$\Phi^w = \frac{7 \times 1.59 - 2 \times 2.59}{\sqrt{(7 \times 2 - 2^2)(7 \times 2.59 - 2.59^2)}} = 0.56 \quad (3)$$

The weighted version balances the reduced terms n_p^w in the numerator of the equation with N_p^w in the numerator and denominator meaning the equation still gives both positively and negatively correlated habitats. If all of the locations within the analysis are certain (one habitat), the weighting is 1 and the

result is the same as the original correlation index. See Appendix 2 for matrix representation of the data and other equations.

We calculated both the original and the uncertainty-weighted correlation index ϕ values and permutated (De Cáceres and Legendre 2009) to get a p-value for each habitat and for each carabid species. See De Cáceres and Legendre (2009) for additional considerations when conducting permutation tests.

Data

We used the large volume of carabid (Coleoptera: Carabidae) location records and high-quality land cover data available in Great Britain.

Carabid data

The National Biodiversity Network (NBN) atlas (2017) contains presence records for many species, at 100 m resolution resulting from the six digit Ordnance Survey grid reference (Telfer 2006). We downloaded all records of carabid locations from the NBN atlas on the 7/11/2017 and initially selected those above an arbitrary threshold of at least 10 records (268 species). We converted the coordinates into 100 m grid cells, with the coordinates representing the bottom left corner, using ArcGIS (v 10.4.1 © 2016 ESRI, Redlands, California). NBN species names were checked and synonyms corrected using the Natural History Museum UK species inventory checklist (Raper 2014). Remaining synonyms were corrected using the checklist in Luff (2007). These steps increased the number of records for species with accepted names on these checklists.

The NBN does not include absence data. A species cannot be considered to be absent from all locations where it is not recorded. To allow us to have confidence that a species was genuinely not at a particular location, we counted the number of other species found in each location as a measure of survey effort. Following Hickling et al (2006) we considered a location to be a true absence if it had more records than a threshold number of other carabids. The threshold number of species is arbitrary. For butterflies, Redhead et al. (2016) used a value of 10% of the species pool (5 species). Using 10% of the carabid species would have required 28 or more species, giving only 94 locations across Britain. We used a threshold of 14 species (5%) giving 556 potential absence grid cells and conducted a sensitivity analysis of this value. Absence locations for a species are the remainder of these 556 grid cells after removing those containing the species of interest.

Land cover

We used the vector LCM2015 for Great Britain (Rowland et al. 2017a) to provide habitat data. LCM2015 contains 21 land cover classes based on the UK Biodiversity Action Plan Broad Habitats (Jackson 2000). These classes are assigned to Ordnance Survey Master Map polygons using a Random Forest object-based classification of satellite Landsat-8 (30 m resolution) and AWIFS data (60 m resolution) (Rowland et al. 2017b). Polygons smaller than 0.5 ha or less than 50 m in width are merged into neighbouring polygons. This can remove linear habitats such as those within freshwater, only capturing larger water bodies and wide rivers (Rowland et al. 2017b).

We intersected LCM2015 data with the 100 m NBN squares and calculated the proportion of each habitat at each location. In principle, one might include temperature or altitude, or group land cover classes. Analysing a large number of alternative habitats can lead to a loss of power. Therefore, if dividing some habitats, others should be amalgamated. But here, for simplicity, we used the LCM2015 classes as they are without further classification.

Validation data

To allow validation of both weighted and original correlation indices we used information from Luff, (2007) "The Carabidae (ground beetles) of Britain and Ireland". Luff (2007) is a comprehensive text on British carabid identification including descriptions of where the species might be found. We used only habitat preferences within the British Isles due to differences in associations to other parts of Europe (Eversham and Telfer 1994, Desender et al. 2005). Luff, (2007) stated the preferred habitat for each carabid species in a descriptive way; for example, "In most habitats, especially agricultural fields, gardens and other disturbed, open and dry situations" (p. 68, Trechus quadristriatus). Luff (2007) did not create the book as a database of species associations. It was, therefore, necessary to convert the text into a database against which we could compare our analyses.

We developed a method using as little subjective interpretation as possible. We looked at all words in the descriptions in Luff (2007) of habitat and picked out those words naming a habitat. We then translated these, into either an individual or group of LCM2015 habitat classes. For example, "moorland" in Luff was translated as including Inland rock (in LCM2015 documentation included under "Mountain, heath,

bog" (Rowland et al. 2017b)), Acid grassland, Heather grassland, and Heather & Bog. Where Luff's habitat descriptions represented a group of land covers, the group was included in the database as an aggregate class against which to check the analysis. For a table showing a full list of the words used and resulting LCM2015 habitat classes and aggregates (see Table S1).

Analyses

The NBN data contained a separate record for each species at each relevant 100 m location meaning that individual locations appeared multiple times. We created a version of the data with each location represented once, giving presence or absence (absences determined as described above) for each species at that location. We created this wide format version by using an R script to go through each location and assign a new binary column of presence for each species. Table 1 shows an example of the data after pre-processing. The correlation index and permutations of the analysis, for each species and versions of the method, were processed using the JASMIN cluster (Lawrence et al. 2013). The R scripts for all analyses are given in Appendix 3.

Sensitivity analysis

We conducted a sensitivity analysis of threshold number of species used to define absence locations by using Spearman's rank correlation to determine to what extent the order of the habitat associations from positive to negative ϕ changed using seven (2.5% of the total species number) and 28 (10%) species number thresholds compared with the baseline of 14 . We also compared the order of habitat associations from positive to negative ϕ between the weighted and original index for each species using Spearman's rank correlation.

Validation

The correlation index results for each carabid species were validated by comparing them to the database created from Luff (2007) (section 2.3). For each species, we calculated the percentage of "Luff habitats" that were also found to be significantly (p-value ≤ 0.05) and positively associated habitats in our correlation analysis for that species.

Results

By allowing the use of locations containing more than one habitat, the weighted index used more records for each species and therefore included 52 extra species; 19% more. For example, for Bembidion prasinum the original method only included 14 records, but the weighted method used 79 records. Luff (2007) describes this species as living in shingle near running water. The original method did not include freshwater at all due to a lack of records. The weighted method associated the species most strongly with freshwater. Comparing the rank of the habitats based on their Phi score for the weighted and original analyses for this species using Spearman's rank, the rho value was only 0.62. The species that have far fewer records in the original than the weighted version, like B. prasinum, drove the average correlation down. In most cases where both species had many records, the rank correlation was higher. One exception to high correlation with many records is Curtonotus aulicus that had 106 original and 258 weighted records. The original version had freshwater non-significantly ($p = 0.392$) positive despite this being described as a dry habitat species (Luff 2007). The weighted analysis of C. aulicus had freshwater as the habitat most significantly ($p = 8.00 \times 10^{-04}$) avoided.

Validation

Using the 14 species threshold for absence, the original version had 207 and the weighted version 264 species with at least one significant habitat association. Furthermore, the weighted and original indices gave similar ranked habitat associations, with the average Spearman's rank correlation 0.82 (SE 0.008) between the two indices. That is not to say however, that significant results sensibly described the habitat of the species. We, therefore validated the correlation results against the database created from Luff (2007).

Considering the average (across species) percentage match of our analyses to Luff habitats, the original analysis identified on average 68% (using 187 species) of Luff habitats and the weighted analysis 70% (using 239 species). This is not a great deal more on average, but does include more species. In the original version, all of the Luff habitats were identified for 94 species and at least one Luff habitat for 157 species. In the weighted analysis, all of the Luff habitats were selected for 126 species and at least one Luff habitat for 205 species. Comparing with Luff (2007), the weighted version matched 18 species less

well than the original version, 141 matched as well, and 28 matched better. Overall, using only the species analysed using both methods, the weighted version matched 6.8% on average better compared to the original version. Fig 1 shows the graphical comparison of the two versions of the index. The weighted version generally gave a slightly higher percentage matches for species with a moderate to large number of records, and included more species with few records.

Individual species examples.

Here we give examples showing comparisons between the original and weighted version of the index, the improvement using the weighted method and establishing how few records are required to give a reasonable estimate of habitat preference. For the full dataset of all carabids analysed see Appendix 4.

Original vs weighted index

Abax parallelepipedus is described by Luff (2007) as a woodland and moorland species. Due to insufficient data, the original version failed to classify three habitats, despite having 176 records, but did show a preference for woodland and heather grassland (Fig 2). The weighted method classified all habitats and captured the woodland and more of the moorland habitat types. For Acupalpus dubius neither analysis matched Luff ("In litter, moss and tussocks near fresh water"(p.175) translated as Freshwater), but may give additional information (Fig 2) as an association was found with "Fen, marsh and swamp", potentially represent the moss and tussocks of Luff's description. The analyses identifies freshwater for other waterside species (see Appendix 4), this therefore is not a consistent problem with detecting freshwater. Two examples are; Anthracus consputus and Trechoblemus micros, which both Luff and our analysis classify as freshwater species.
Calathus fuscipes and Loricera pilicornis are two examples of species that matched Luff habitats better in the original than the weighted version, which failed to match open grassland and suburban respectively (Fig 3). For both species the named habitat remained positively associated in the weighted analysis, but had higher p-values, 0.16 and 0.23 respectively.

Number of records required

Species with between 10 and 35 records in the weighted analysis gave matches with an average of 66% of Luff habitats. With so few presence records, however, the analysis had less power to differentiate habitats and to detect significance. For Amara curta the analyses was not able to detect any avoided habitats and analysis failed to pick up on the heath association suggested by Luff (2007). With 50 or 60 records, as in the case of Bracteon litorale or Harpalus anxius, the analysis was more able to differentiate the individual habitats. Bracteon litorale, which Luff (2007) describes as "On bare sand and fine shingle near rivers or standing water", was associated in our analysis with broadleaved woodland and improved grassland, as well as agreeing with Luff by including freshwater. For Harpalus anxius, the analysis seemed to select the dunes of Luff's description well, with supralittoral sediments the most preferred habitat, but did not select heaths. Additionally a positive association with saltmarsh was identified, which is often near dunes (Fig 4).

Sensitivity analysis

Spearman's rank correlation values were high when comparing habitat association calculated with the threshold value of 14 to a threshold of seven or 28 (Table 2). Even comparing the seven to the 28 threshold, the rank of the habitats remained consistent.

Discussion

Our new weighted version of the Phi correlation index allowed substantially more records to be included for each species and therefore increased the number of species that could be analysed and improved the predictions of habitat association. The use of the number of species records as a proxy of survey effort was robust, being insensitive to the threshold for defining absence locations. The weighted analysis was able to give accurate results with as few as 50 records, and the use of absences enhanced the ability to determine habitat associations. Informative results using so few records are in stark contrast to other methods which require thousands of records for each species. Redhead et al. (2016) suggest that few taxa are well-enough recorded to provide so many records, our improved method will be applicable to many more taxa. For example, 35% of cerambycid beetles have 50 or more records in Great Britain (44% for carabids). Our method also gives a target for recording the rarer specialist species, whose conservation most requires an evidence base (Lawton et al. 2010).

As the number of records gets very large the Phi coefficient becomes the Ochiai index, which is itself related to a modified version of IndVal (De Cáceres et al. 2008). The number of records in the data we have are not large and the Ochiai index was therefore not applicable. It is possible, however, to extend both the non-equalized and group-equalised IndVal in a similar way to the phi coefficient we present in this paper by adding habitat weighting. The values still range between zero and one and the weighted version gives a value for more of the habitats. The results of weighting IndVal have not been tested, but this could be done in future research identifying indicator species. To facilitate such a test, this capability is included in our PhiCor R package. Dufrêne and Legendre (1997) used carabid data from pitfall traps to validate IndVal originally. The capture locations of all individuals were known precisely. However, besides using the weighting for imprecise locations, as presented in our case study on the phi coefficient, the weighting method could be useful in cases with precise locations for a number of the indices presented in De Cáceres and Legendre (2009). These cases include species foraging or dispersing into neighbouring habitats (McIntire et al. 2013), source-sink dynamics of plants (Kadmon and Shmida 1990), or to account for the uncertainty of land cover classification (Morton et al. 2011). It may even be worthwhile drawing buffers around record locations so as to include information on surrounding habitat.
Unlike species distribution models, the correlation index does not suffer from overfitting (Breiner et al. 2015). However, as numbers of presences and absences differ between species, comparison among species is not straightforward. The maximum ϕ values vary with the number of records and are rarely comparable between species. The rank of the habitats is comparable but where two species have similar ranks for a habitat they may not have the same affinity. The number of positive habitats for each species, however, is positively correlated with the degree to which a species is categorised independently as generalist vs specialist (see Appendix 5).
One possible way of increasing the comparability between species is to use the group equalised correlation index (Tichy and Chytry 2006). Beyer et al. (2010) reviewed the factors influencing habitat preference of species, arguing that species which are found more often are so because the habitat is more common. Tichy & Chytry (2006) suggested a group (habitat) equalised version of the correlation index. This version modifies several of the inputs by the number of groups. In our case, group equalising usually resulted in the same habitats having significant associations, although the ϕ values were often different. As an example, Bembidion lampros is associated in the non-equalised analysis with arable followed by conifer and urban. In the equalised analysis the same habitats are retained in the top three, but now the beetle is most associated with coniferous, urban and then arable. A weighted group-equalised version (Appendix 6) did not match the Luff (2007) validation data quite as well, but is included in the full output (Appendix 4). It should be noted that species may not be equally detectable in different habitats and therefore, where the data is available a similar equalisation could be done using detectability.
The analysis we have conducted agrees to some extent with previous smaller scale studies of carabids using different analytical approaches. Eyre and Luff (2004) used constrained ordination with 126 carabid species against the proportion of 12 habitats within 1 km squares across north-east England and south-east Scotland. Some of their results agree with ours, although, as an example, their analysis suggests a higher preference of Abax parallelepipedus for inland water than broadleaved woodland. Eyre and Luff (2004) point out that some unexpected relationships of species and land covers suggest care is needed when interpreting their results and that the low eigenvalues and cumulative percentage variation suggest noisy data.
Within the literature the same species is sometimes attributed to different habitats in different studies without clear information on where this association information stems from or the species' other associations. An example is Pterostichus madidus, which is variously described as inhabiting dry open, urban, moorland or grassland (Butterfield et al. 1995, Dennis et al. 2004, Angold et al. 2006, Morecroft et al. 2009), with Luff (2007) describing the species as "woodland, garden and dry grassland". Our analysis agrees with all of these habitats, suggesting the species is associated with a wide range of habitats. The method we present provides a robust method of presenting all the associations of a species, which can be used to paint a clearer picture of habitat associations.
We chose in the main analysis to remove record locations with more than one habitat. Another option was to choose the most abundant habitat. We conducted a version of the unweighted analysis choosing the most abundant habitat in each 100 m square. This version matched the Luff (2007) validation less well than the unweighted version removing records (Appendix 1). This is likely due to misclassification of the habitat that the species was found in or the loss of information about which habitat individuals of the species could have been in prior to being caught.
In conclusion, our new weighted method demonstrates an improvement to the Phi coefficient of association, which is simpler than ordination, requires fewer records than regression, and gives habitat preference and avoidance. Our method allows for uncertainty in the habitats associated with the record location and is ideal for mobile species, which may be found outside of preferred habitats. It utilises more

of existing sources of data, including every habitat within a non-point location, giving quantitative information on habitat preference. Our work provides guidance on the flexible threshold defining absence records and targets for the number of records necessary to achieve a reasonable result for each species. The method is usable as-is to provide detailed data usable in conservation planning and the case study provides the carabid analysis ready to use in modelling and improving interpretation of the results of future studies. Having established the method as working for carabids, the method would benefit from further testing with different taxa.

Acknowledgements

This work used the JASMIN at RAL STFC (http://jasmin.ac.uk), operated jointly by the centre of environmental data analysis and the scientific computing department. This facility was funded by NERC. We thank all of the individual contributors to the NBN atlas data (Appendix 7)

Data Accessibility

CEH LCM2015 (Rowland et al. 2017a) is available for academic purposes from https://doi.org/10.5285/6c6c9203-7333-4d96-88ab-78925e7a4e73. National Biodiversity Network (NBN) atlas (2017) carabid data is available from https://species.nbnatlas.org/species/NHMSYS0001717497.

Author specific acknowledgements

Jordan Chetcuti was funded by a studentship from the NERC SPHERES Doctoral Training Partnership (NE/L002574/1).

Author specific Data Accessibility

R scripts used in the analysis can be found in the supplementary information and outputs on http://eidc.ceh.ac.uk. (This gives a clue that it's someone at CEH). Additionally we have published an R package to allow others to use the method (https://github.com/Zabados/PhiCor) and created a webpage to allow all of the results from the case study to be visualised and compared (https://shiny-apps.ceh.ac.uk/CarabidData/).

References

Angold, P. G. et al. 2006. Biodiversity in urban habitat patches. - Sci. Total Environ. 360: 196–204.

Ball, O. J.-P. et al. 2013. Habitat associations and detectability of the endemic Te Paki ground beetle Mecodema tenaki (Coleoptera: Carabidae). - N. Z. J. Ecol. 37: 84–94.

Beyer, H. L. et al. 2010. The interpretation of habitat preference metrics under use-availability designs. - Philos. Trans. R. Soc. B Biol. Sci. 365: 2245–2254.

Breiner, F. T. et al. 2015. Overcoming limitations of modelling rare species by using ensembles of small models. - Methods Ecol. Evol. 6: 1210–1218.

Brodie, J. F. et al. 2016. Connecting science, policy, and implementation for landscape-scale habitat connectivity. - Conserv. Biol. 30: 950–961.

Butterfield, J. et al. 1995. Carabid beetle communities as indicators of conservation potential in upland forests. - For. Ecol. Manage. 79: 63–77.

Cole, L. J. et al. 2010. The influence of fine-scale habitat heterogeneity on invertebrate assemblage structure in upland semi-natural grassland. - Agric. Ecosyst. Environ. 136: 69–80.

De Cáceres, M. and Legendre, P. 2009. Associations between species and groups of sites: indices and statistical inference. - Ecology 90: 3566–3574.

De Cáceres, M. et al. 2008. Assessing species diagnostic value in large data sets: A comparison between phi-coefficient and Ochiai index. - J. Veg. Sci. 19: 779–788.

De Gasperis, S. R. et al. 2016. Distribution and abundance of hole-nesting birds in Mediterranean forests: Impact of past management patterns on habitat preference. - Ornis Fenn. 93: 100–110.

De Lima, R. F. et al. 2016. Distribution and habitat associations of the critically endangered bird species of São Tomé Island (Gulf of Guinea). - Bird Conserv. Int.: 1–15.

Dennis, P. et al. 2004. Consequences for biodiversity of reducing inputs to upland temperate pastures: effects on beetles (Coleoptera) of cessation of nitrogen fertilizer application and reductions in stocking rates of sheep. - Grass Forage Sci. 59: 121–135.
Desender, K. et al. 2005. Rural-urban gradients and the population genetic structure of woodland ground beetles. - Conserv. Genet. 6: 51–62.
Dufrêne, M. and Legendre, P. 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. - Ecol. Monogr. 67: 345–366.
Eversham, B. C. and Telfer, M. G. 1994. Conservation value of roadside verges for stenotopic heathland Carabidae: corridors or refugia? - Biodivers. Conserv. 3: 538–545.
Eyre, M. D. and Luff, M. L. 2004. Ground beetle species (Coleoptera, Carabidae) associations with land cover variables in northern England and southern Scotland. - Ecography (Cop.). 27: 417–426.
Ferrão, M. et al. 2018. A new species of Amazonian snouted treefrog (Hylidae: Scinax) with description of a novel species-habitat association for an aquatic breeding frog. - PeerJ 6: 34.
Gardner, S. M. 1991. Ground Beetle (Coleoptera : Carabidae) Communities on Upland Heath and Their Association with Heathland Flora. - J. Biogeogr. 18: 281–289.
Hickling, R. et al. 2006. The distributions of a wide range of taxonomic groups are expanding polewards. - Glob. Chang. Biol. 12: 450–455.
Hill, M. O. et al. 1975. Indicator Species Analysis , A Divisive Polythetic Method of Classification , and its Application to a Survey of Native Pinewoods in Scotland. - J. Ecol. 63: 597–613.
Jackson, D. L. 2000. Guidance on the interpretation of the Biodiversity Broad Habitat Classification (terrestrial and freshwater types): Definitions and the relationship with other classifications, JNCC Report 307.
Kadmon, R. and Shmida, A. 1990. Spatiotemporal demographic processes in plant populations: An approach and a case study. - Am. Nat. 135: 382.
Lawrence, B. N. et al. 2013. Storing and manipulating environmental big data with JASMIN. - Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 6-9 2013: 68–75.
Lawton, J. H. et al. 2010. Making space for nature: A review of England's wildlife Sites and ecological network. - Rep. to Defra: 107.
Lonsdorf, E. et al. 2009. Modelling pollination services across agricultural landscapes. - Ann. Bot. 103: 1589–1600.
Luff, M. L. 2007. The Carabidae (ground beetles) of Britain and Ireland. - Royal Entomological Society.
Maclean, I. M. D. et al. 2011. Predicting changes in the abundance of African wetland birds by incorporating abundance-occupancy relationships into habitat association models. - Divers. Distrib. 17: 480–490.
McIntire, E. J. B. et al. 2013. Biased correlated random walk and foray loop: Which movement hypothesis drives a butterfly metapopulation? - Oecologia 172: 293–305.
Morecroft, M. D. et al. 2009. The UK Environmental Change Network: Emerging trends in the composition of plant and animal communities and the physical environment. - Biol. Conserv. 142: 2814–2832.
Morton, D. et al. 2011. Final Report for LCM2007 - the new UK land cover map. Countryside Survey Technical Report No 11/07 NERC/Centre for Ecology & Hydrology 112pp. (CEH Project Number: C03259).
NBN Atlas website 2017. Carabidae : Ground beetle at https://species.nbnatlas.org/species/NHMSYS0001717497. Accessed 11 November 2017.
NBN Atlas website 2018. at https://nbnatlas.org. Accessed 13 April 2018.
Pearson, K. 1896. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. - Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 187: 253–318.
Petit, S. et al. 2003. Knowledge-based models for predicting species occurrence in arable conditions. - Ecography (Cop.). 26: 626–640.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. - Ecol. Modell. 190: 231–259.
Porto, T. J. et al. 2018. Regional distribution patterns can predict the local habitat specialization of arachnids in heterogeneous landscapes of the Atlantic Forest. - Divers. Distrib. 24: 375–386.
Pouzols, F. M. and Moilanen, A. 2014. A method for building corridors in spatial conservation prioritization. - Landsc. Ecol. 29: 789–801.
Raper, C. 2014. Dataset: UK Species Inventory. Natural History Museum Data Portal DOI: 10.5519/0038741 Accessed 15 January 2018.
Redhead, J. W. et al. 2016. Assessing species' habitat associations from occurrence records, standardised monitoring data and expert opinion: A test with British butterflies. - Ecol. Indic. 62: 271–278.
Ricotta, C. et al. 2015. Let the concept of indicator species be functional! - J. Veg. Sci. 26: 839–847.

Ries, L. et al. 2004. Ecological Responses to Habitat Edges: Mechanisms, Models, and Variability Explained. - Annu. Rev. Ecol. Evol. Syst. 35: 491–522.
Rowland, C. S. et al. 2017a. Land Cover Map 2015 (vector, GB). - NERC Environ. Inf. Data Cent.
Rowland, C. S. et al. 2017b. Land Cover Map 2015 Dataset documentation. v1.2
Telfer, M. G. 2006. Ground Beetle Recording Scheme.
Tichy, L. and Chytry, M. 2006. Statistical determination of diagnostic species for site groups of unequal size. - J. Veg. Sci. 17: 809–818.
Webb, J. et al. 2017. Pantheon - database version 3.7.4.

Figure Legends

Fig 1 Comparison of original and weighted correlation index showing how they match the validation data. Species are in bins of the number of records (using the records without removal, as used in the weighted version). Species that did not have enough records in the original version are included on the left to show that the weighted version on average when including these species achieves a match with the validation data.

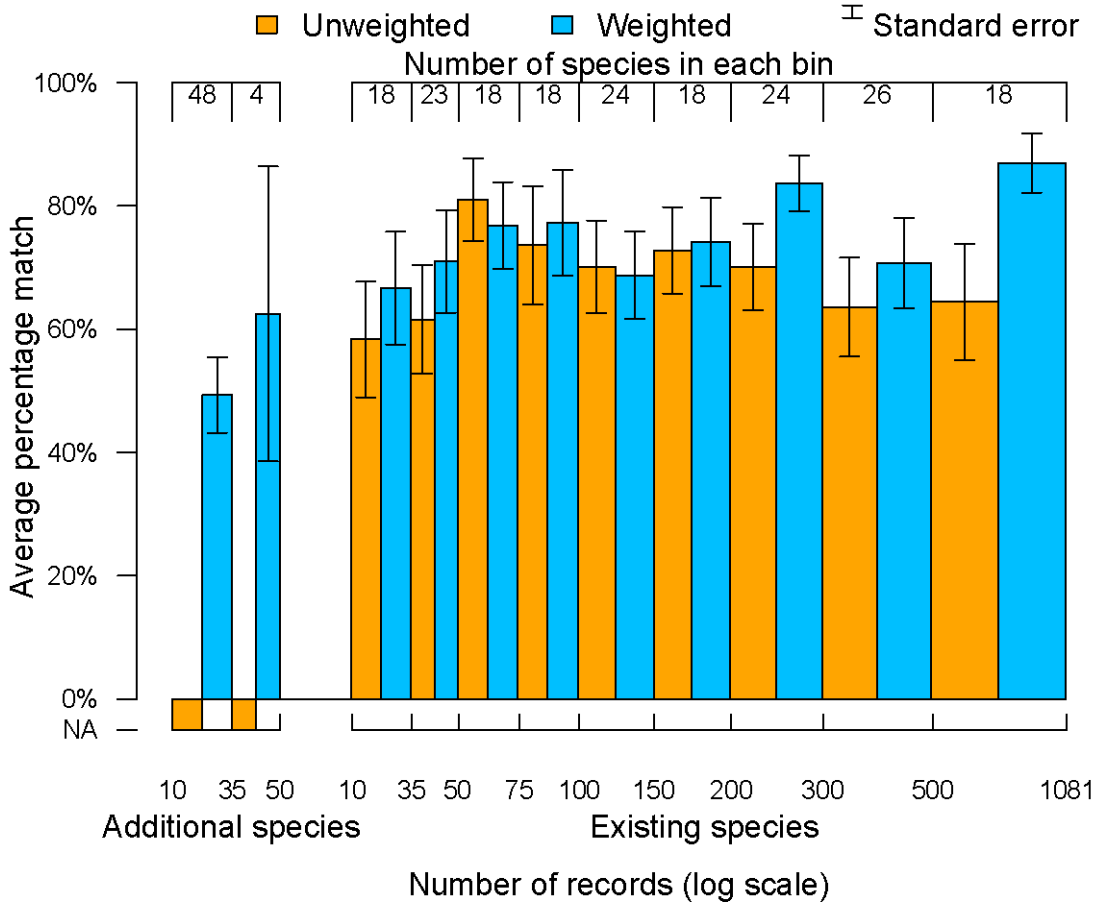
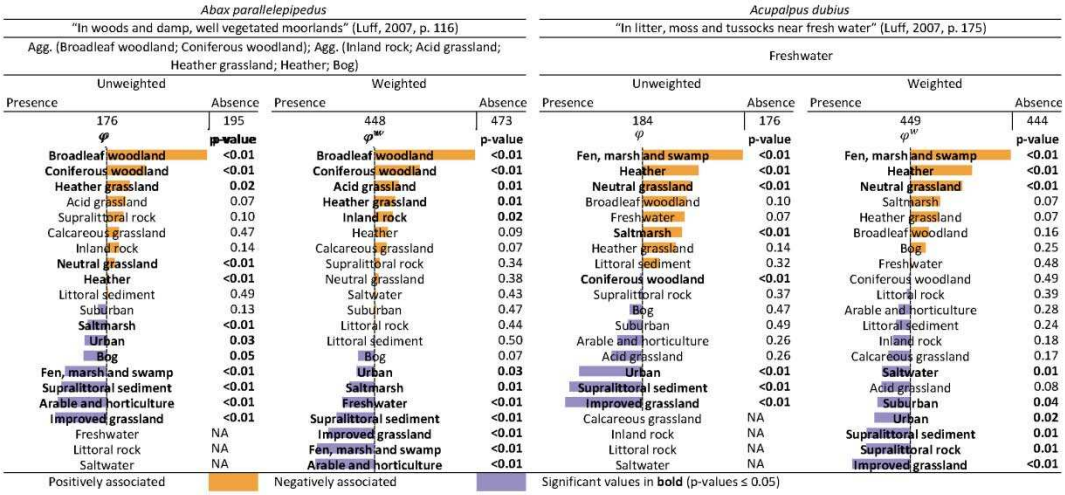


Fig 2 Abax parallelepipedus and Acupalpus dubius original and weighted habitat correlation analysis showing the relative positive and negative ϕ and p values. These examples show the improvement offered using the weighted method, matching better with Luff and including more habitats.



Abax parallelepipedus

"In woods and damp, well vegetated moorlands" (Luff, 2007, p. 116)

Agg. (Broadleaf woodland; Coniferous woodland); Agg. (Inland rock; Acid grassland; Heather grassland; Heather; Bog)

Unweighted			Weighted		
Presence		Absence	Presence		Absence
176		195	448		473
φ		p-value	φ^w		p-value
Broadleaf woodland		**<0.01**	**Broadleaf woodland**		**<0.01**
Coniferous woodland		**<0.01**	**Coniferous woodland**		**<0.01**
Heather grassland		**0.02**	**Acid grassland**		**0.01**
Acid grassland		0.07	**Heather grassland**		**0.01**
Supralittoral rock		0.10	**Inland rock**		**0.02**
Calcareous grassland		0.47	Heather		0.09
Inland rock		0.14	Calcareous grassland		0.07
Neutral grassland		**<0.01**	Supralittoral rock		0.34
Heather		**<0.01**	Neutral grassland		0.38
Littoral sediment		0.49	Saltwater		0.43
Suburban		0.13	Suburban		0.47
Saltmarsh		**<0.01**	Littoral rock		0.44
Urban		**0.03**	Littoral sediment		0.50
Bog		**0.05**	**Bog**		0.07
Fen, marsh and swamp		**<0.01**	**Urban**		**0.03**
Supralittoral sediment		**<0.01**	**Saltmarsh**		**0.01**
Arable and horticulture		**<0.01**	**Freshwater**		**<0.01**
Improved grassland		**<0.01**	**Supralittoral sediment**		**<0.01**
Freshwater		NA	**Improved grassland**		**<0.01**
Littoral rock		NA	**Fen, marsh and swamp**		**<0.01**
Saltwater		NA	**Arable and horticulture**		**<0.01**
Positively associated			Negatively associated		

Acupalpus dubius

"In litter, moss and tussocks near fresh water" (Luff, 2007, p. 175)

Freshwater

Unweighted			Weighted		
Presence		Absence	Presence		Absence
184		176	449		444
φ		p-value	φ^w		p-value
Fen, marsh and swamp		**<0.01**	**Fen, marsh and swamp**		**<0.01**
Heather		**<0.01**	**Heather**		**<0.01**
Neutral grassland		**<0.01**	**Neutral grassland**		**<0.01**
Broadleaf woodland		0.10	Saltmarsh		0.07
Freshwater		0.07	Heather grassland		0.07
Saltmarsh		**<0.01**	Broadleaf woodland		0.16
Heather grassland		0.14	Bog		0.25
Littoral sediment		0.32	Freshwater		0.48
Coniferous woodland		**<0.01**	Coniferous woodland		0.49
Supralittoral rock		0.37	Littoral rock		0.39
Bog		0.47	Arable and horticulture		0.28
Suburban		0.49	Littoral sediment		0.24
Arable and horticulture		0.26	Inland rock		0.18
Acid grassland		0.26	Calcareous grassland		0.17
Urban		**<0.01**	**Saltwater**		**0.01**
Supralittoral sediment		**<0.01**	Acid grassland		0.08
Improved grassland		**<0.01**	**Suburban**		**0.04**
Calcareous grassland		NA	**Urban**		**0.02**
Inland rock		NA	**Supralittoral sediment**		**0.01**
Littoral rock		NA	**Supralittoral rock**		**0.01**
Saltwater		NA	**Improved grassland**		**<0.01**

Significant values in **bold** (p-values ≤ 0.05)

Fig 3 Calathus fuscipes and Loricera pilicornis original and weighted habitat correlation analysis showing the relative positive and negative ϕ and p values. Showing that in these cases the original version matched more Luff habitats than the weighted version. The weighted method does however manage to represent more of the habitats.

Calathus fuscipes

"In open grasslands, arable fields and gardens" (Luff, 2007, p. 121)

Agg. (Acid grassland; Calcareous grassland; Heather grassland; Neutral grassland; Improved grassland); Agg.(Arable and horticulture; Improved grassland); Agg.(Suburban; Urban)

Unweighted			Weighted		
Presence		Absence	Presence		Absence
163		138	394		404
φ		p-value	φ^w		p-value
Supralittoral sediment		**<0.01**	**Supralittoral sediment**		**<0.01**
Urban		**<0.01**	**Littoral sediment**		**0.00**
Inland rock		0.16	**Urban**		**0.04**
Acid grassland		0.30	Acid grassland		0.06
Heather grassland		**<0.01**	Inland rock		0.09
Littoral sediment		0.31	Calcareous grassland		0.07
Coniferous woodland		**<0.01**	Littoral rock		0.08
Suburban		0.44	Heather grassland		0.16
Neutral grassland		0.70	Heather		0.20
Improved grassland		0.41	Coniferous woodland		0.44
Arable and horticulture		0.43	Saltwater		0.43
Heather		0.36	Neutral grassland		0.43
Saltmarsh		**<0.01**	Saltmarsh		0.44
Supralittoral rock		**<0.01**	Suburban		0.21
Broadleaf woodland		**0.02**	Arable and horticulture		0.19
Bog		**<0.01**	Supralittoral rock		0.08
Fen, marsh and swamp		**<0.01**	Improved grassland		0.06
Calcareous grassland		NA	**Bog**		**0.01**
Freshwater		NA	**Broadleaf woodland**		**<0.01**
Littoral rock		NA	**Freshwater**		**<0.01**
Saltwater		NA	**Fen, marsh and swamp**		**<0.01**

Positively associated — Negatively associated

Loricera pilicornis

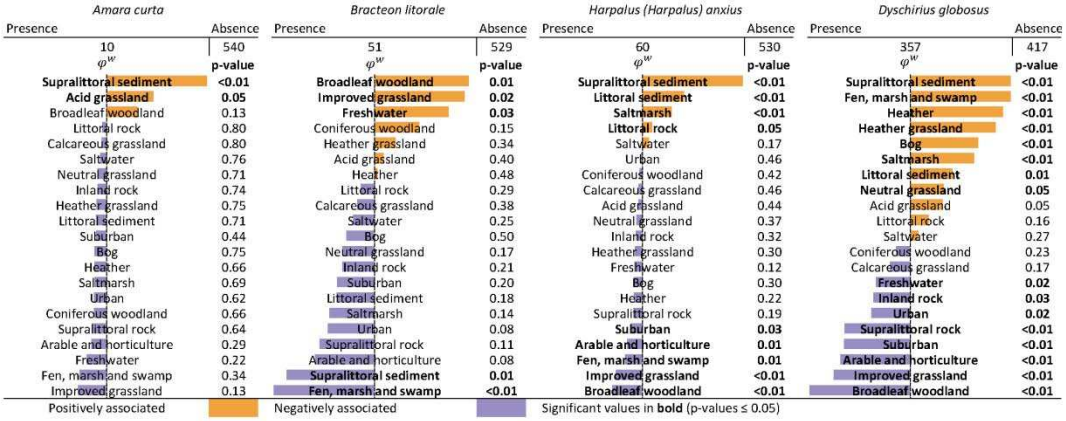
"In grasslands, damp woodland, cultivated fields, gardens and near standing or running fresh water" (Luff, 2007, p. 56)

Freshwater; Agg. (Acid grassland; Calcareous grassland; Heather grassland; Neutral grassland; Improved grassland); Agg. (Broadleaf woodland; Coniferous woodland); Agg.(Arable and horticulture; Improved grassland); Agg.(Suburban; Urban)

Unweighted			Weighted		
Presence		Absence	Presence		Absence
375		107	913		281
φ		p-value	φ^w		p-value
Bog		**0.01**	**Bog**		**<0.01**
Broadleaf woodland		**0.02**	**Heather**		**<0.01**
Heather		**0.02**	**Broadleaf woodland**		**0.01**
Acid grassland		0.10	**Acid grassland**		**0.02**
Coniferous woodland		**<0.01**	Fen, marsh and swamp		0.07
Neutral grassland		**<0.01**	**Coniferous woodland**		**0.04**
Suburban		**<0.01**	Heather grassland		0.21
Inland rock		0.30	Suburban		0.23
Fen, marsh and swamp		0.43	Supralittoral sediment		0.25
Heather grassland		**<0.01**	Saltmarsh		0.42
Supralittoral sediment		0.36	Neutral grassland		0.47
Saltmarsh		**<0.01**	Littoral rock		0.46
Urban		**<0.01**	Calcareous grassland		0.32
Improved grassland		**0.03**	Freshwater		0.24
Littoral rock		**0.04**	Inland rock		0.11
Supralittoral rock		**0.01**	Saltwater		0.08
Arable and horticulture		**<0.01**	Urban		0.07
Calcareous grassland		NA	**Littoral sediment**		**0.01**
Freshwater		NA	**Improved grassland**		**<0.01**
Littoral sediment		NA	**Arable and horticulture**		**<0.01**
Saltwater		NA	**Supralittoral rock**		**<0.01**

Significant values in **bold** (p-values ≤ 0.05)

Fig 4 Amara curta, Bracteon litorale, Harpalus anxius, and Dyschirius globosus weighted habitat correlation analysis showing the relative positive and negative ϕ and p values. Showing that with more than 50 records the analysis gives both significantly positive and negative association.



Amara curta

Presence		Absence
10		540
	φ^w	p-value
Supralittoral sediment		**<0.01**
Acid grassland		**0.05**
Broadleaf woodland		0.13
Littoral rock		0.80
Calcareous grassland		0.80
Saltwater		0.76
Neutral grassland		0.71
Inland rock		0.74
Heather grassland		0.75
Littoral sediment		0.71
Suburban		0.44
Bog		0.75
Heather		0.66
Saltmarsh		0.69
Urban		0.62
Coniferous woodland		0.66
Supralittoral rock		0.64
Arable and horticulture		0.29
Freshwater		0.22
Fen, marsh and swamp		0.34
Improved grassland		0.13

Bracteon litorale

Presence		Absence
51		529
	φ^w	p-value
Broadleaf woodland		**0.01**
Improved grassland		**0.02**
Freshwater		**0.03**
Coniferous woodland		0.15
Heather grassland		0.34
Acid grassland		0.40
Heather		0.48
Littoral rock		0.29
Calcareous grassland		0.38
Saltwater		0.25
Bog		0.50
Neutral grassland		0.17
Inland rock		0.21
Suburban		0.20
Littoral sediment		0.18
Saltmarsh		0.14
Urban		0.08
Supralittoral rock		0.11
Arable and horticulture		0.08
Supralittoral sediment		**0.01**
Fen, marsh and swamp		**<0.01**

Harpalus (Harpalus) anxius

Presence		Absence
60		530
	φ^w	p-value
Supralittoral sediment		**<0.01**
Littoral sediment		**<0.01**
Saltmarsh		**<0.01**
Littoral rock		**0.05**
Saltwater		0.17
Urban		0.46
Coniferous woodland		0.42
Calcareous grassland		0.46
Acid grassland		0.44
Neutral grassland		0.37
Inland rock		0.32
Heather grassland		0.30
Freshwater		0.12
Bog		0.30
Heather		0.22
Supralittoral rock		0.19
Suburban		**0.03**
Arable and horticulture		**0.01**
Fen, marsh and swamp		**0.01**
Improved grassland		**<0.01**
Broadleaf woodland		**<0.01**

Dyschirius globosus

Presence		Absence
357		417
	φ^w	p-value
Supralittoral sediment		**<0.01**
Fen, marsh and swamp		**<0.01**
Heather		**<0.01**
Heather grassland		**<0.01**
Bog		**<0.01**
Saltmarsh		**<0.01**
Littoral sediment		**0.01**
Neutral grassland		**0.05**
Acid grassland		0.05
Littoral rock		0.16
Saltwater		0.27
Coniferous woodland		0.23
Calcareous grassland		0.17
Freshwater		**0.02**
Inland rock		**0.03**
Urban		**0.02**
Supralittoral rock		**<0.01**
Suburban		**<0.01**
Arable and horticulture		**<0.01**
Improved grassland		**<0.01**
Broadleaf woodland		**<0.01**

Positively associated | Negatively associated | Significant values in **bold** (p-values ≤ 0.05)

Table Legends

Table 1 Example of the vectors that can be used in calculating the Phi coefficient for each individual habitat, showing the , proportion of each habitat within each location and the binary presence data, in this case for the species Abax parallelepipedus (see Appendix 2 for matrix version of this information and equations).

Location ID	**LCM2015 habitat**	**Heather habitat vector (h)**	**Weight vector (w)**	**Species vector (s)**
1	Heather grassland	0	1.00	0
2	Supralittoral sediment	0	1.00	0
3	Heather grassland	0	1.00	0
4	Acid grassland	0	0.02	1
4	Inland rock	0	0.59	1
4	Heather	1	0.39	1
5	Inland rock	0	1.00	0
6	Heather	1	0.76	0
6	Improved grassland	0	0.24	0
7	Inland rock	0	1.00	1
…	…	⋮	⋮	⋮

All three vectors have length L

Table 2 Comparison of the habitat associations using Spearman's rank correlation between different thresholds of species numbers used to define absence squares in the analysis of carabid land cover association.

	Threshold 7 and 14	**Threshold 14 and 28**	**Threshold 7 and 28**	**Number of species**
Original	0.90 (SE 0.004)	0.86 (SE 0.009)	0.80 (SE 0.011)	212
Weighted	0.95 (SE 0.002)	0.89 (SE 0.007)	0.84 (SE 0.009)	268