

Assessing species' habitat associations from occurrence records, standardised monitoring data and expert opinion: a test with British butterflies

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Redhead, J. W., Fox, R., Brereton, T. and Oliver, T. H. ORCID: <https://orcid.org/0000-0002-4169-7313> (2016) Assessing species' habitat associations from occurrence records, standardised monitoring data and expert opinion: a test with British butterflies. *Ecological Indicators*, 62. pp. 271-278. ISSN 1470-160X doi: <https://doi.org/10.1016/j.ecolind.2015.11.004> Available at <https://centaur.reading.ac.uk/57272/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1016/j.ecolind.2015.11.004>

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

1 **Assessing species' habitat associations from occurrence records, standardised monitoring**
2 **data and expert opinion: A test with British butterflies**

3 Redhead, J.W. ^{a,c} (johdhe@ceh.ac.uk)

4 Fox, R. ^b (rfox@butterfly-conservation.org)

5 Brereton T. ^b (tbrereton@butterfly-conservation.org)

6 Oliver, T.H. ^{ac} (t.oliver@reading.ac.uk)

7

8 ^a NERC Centre for Ecology and Hydrology, Maclean Building, Wallingford, Oxfordshire, OX10 8BB, UK.

9 ^b Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset, BH20 5QP, UK.

10 ^c School of Biological Sciences, Harborne Building, University of Reading, Reading, Berkshire, RG6 6AS, UK

11 ^d Corresponding author: Tel: (+44) 1491 692538, Fax: (+44) 1491 692424

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26 **Abstract**

27 Accurate knowledge of species' habitat associations is important for conservation planning and
28 policy. Assessing habitat associations is a vital precursor to selecting appropriate indicator species
29 for prioritizing sites for conservation or assessing trends in habitat quality. However, much existing
30 knowledge is based on qualitative expert opinion or local scale studies, and may not remain accurate
31 across different spatial scales or geographic locations. Data from biological recording schemes have
32 the potential to provide objective measures of habitat association, with the ability to account for
33 spatial variation. We used data on 50 species of British butterfly as a test case to investigate the
34 correspondence of data-derived measures of habitat association with expert opinion, from two
35 different butterfly recording schemes. One scheme collected large quantities of occurrence data (c.3
36 million records) and the other, lower quantities of standardised monitoring data (c.1400 sites). We
37 used general linear mixed effects models to derive scores of association with broad-leaf woodland
38 for both datasets and compared them with scores canvassed from experts.

39 Scores derived from occurrence and abundance data both showed strongly positive correlations
40 with expert opinion. However, only for occurrence data did these fell within the range of
41 correlations between experts. Data-derived scores showed regional spatial variation in the strength
42 of butterfly associations with broad-leaf woodland, with a significant latitudinal trend in 26% of
43 species. Sub-sampling of the data suggested a mean sample size of 5000 occurrence records per
44 species to gain an accurate estimation of habitat association, although habitat specialists are likely to
45 be readily detected using several hundred records. Occurrence data from recording schemes can
46 thus provide easily obtained, objective, quantitative measures of habitat association.

47 **Key words:** spatial variation, recording scheme, citizen science, latitudinal gradient, biological
48 indicators

49 **1. Introduction**

50 Associations between species and habitats are one of the basic principles of ecology (Aarts et al.
51 2013; Yapp 1922). As habitat loss remains the primary cause of global biodiversity declines (Brooks
52 et al. 2006; Thomas et al. 2004) identifying such associations accurately is important for
53 conservation planning, policy and research. Where species are in decline, accurate information on
54 habitat associations is required so that investigations into likely causes, and subsequent
55 implementation of conservation efforts, can be targeted correctly. Likewise, if a particular habitat is
56 undergoing change, well characterised associations enable predications to be made about which
57 species are most likely to be affected. Accurate knowledge of associations is also vital to selecting
58 appropriate indicator species for use in prioritizing sites for conservation, monitoring environmental
59 conditions or assessment of habitat quality (Carignan and Villard 2002).

60 Although the habitat associations of some taxa are well characterised, most species are poorly
61 studied. Even for well-studied taxa there may be limitations to our understanding of habitat
62 associations at large spatial scales (Gregory and Baillie 1998) as many studies are carried out at a
63 local level in response to specific conservation issues (e.g. Knight and Arthington 2008; Loeb et al.
64 2000; Rouquette and Thompson 2005). As a result, information on wider scale habitat associations,
65 including that which forms the foundations of much conservation policy, is often extrapolated from
66 such studies or from qualitative descriptions based on expert opinion (Reif et al. 2010). This is
67 potentially problematic, as both habitat associations and expert perceptions of them have been
68 demonstrated to vary with location (O'Leary et al. 2009; Oliver et al. 2009), spatial scale (Mayor et al.
69 2009) and environmental change (Pateman et al. 2012). It is thus important to test existing
70 knowledge on habitat associations against quantitative methods. These have the potential to
71 operate at a range of spatial scales, and to take into account spatial or temporal variation. Such
72 methods also have the potential to uncover cryptic requirements or previously unknown plasticities
73 in habitat association.

74 National or international biological recording and monitoring schemes provide a valuable source of
75 data for analysing large scale patterns in time and space (Bishop et al. 2013; Thomas 2005). Large
76 sample sizes and extensive spatial coverage make them well suited to use in detecting habitat
77 associations. However, monitoring scheme data vary in quality and quantity, from simple
78 occurrence data (i.e. georeferenced records of species' presence) to detailed demographic data from
79 standardised protocols. Whilst datasets at all points along this spectrum have their value for specific
80 applications, it is important to test which are most suitable for detecting habitat associations,
81 especially as increasing levels of information come at a cost of time and effort in collection, and,
82 consequently, in the number and spatial coverage of records (Bishop et al. 2013).

83 This study used two different butterfly recording scheme datasets - one comprising large quantities
84 of occurrence data and the other, lower quantities of abundance data from a standardised
85 monitoring scheme - alongside data on the extent of British broad-leaf woodland. Butterflies are a
86 useful test case for determining habitat associations. They are frequently used as indicator species
87 (Thomas 2005) as their host plant specificity and temperature-dependent development and
88 behaviour make them sensitive to environmental changes, whilst their short life cycles ensure that
89 they respond quickly (Oliver et al. 2009; Pateman et al. 2012; Warren et al. 2001). In Britain, they
90 are well recorded, giving sufficient data for analyses, and well-studied, such that expert opinions are
91 likely to be well-founded and consistent and thus a good yardstick by which to measure the
92 performance of data-derived measures of habitat association. We compared data-derived methods
93 for calculating metrics of habitat association from the two butterfly datasets with expert opinion,
94 including their ability to account for spatial variation in association, and assessed the applicability of
95 these methods to other taxa for which data-derived methods might form the only means by which
96 to assess species' habitat associations.

97 **2. Methods**

98 2.1. SPECIES DATA

99 We obtained data on 50 butterfly species in Great Britain (GB) from two monitoring schemes –
100 Butterflies for the New Millennium (BNM) and the UK Butterfly Monitoring Scheme (UKBMS).
101 Species nomenclature follows Agassiz et al. (2013).

102 BNM is a national scheme which collates butterfly records (i.e. species occurrence at a location),
103 with the aim of maintaining an up-to-date database of butterfly distributions (Asher et al. 2001).
104 This study included only BNM records with spatial resolution of 1 km x 1 km Ordnance Survey grid
105 cell or finer. Duplicate records of the same species in the same cell were removed, resulting in a
106 dataset of approximately 3 million butterfly occurrence records. The study used records from 1990 -
107 2010, to decrease the likely effect of changes over time in woodland extent or habitat association on
108 the results.

109 The UKBMS differs from BNM in aiming to monitor population trends through a standardized survey
110 method involving weekly visits between April and September (Pollard and Yates 1993). Although
111 this allows calculation of abundance throughout each survey year and thus analysis of population
112 trends and phenology, it is relatively labour intensive and there are records from far fewer sites than
113 in BNM (data from 1433 sites were included in our analysis).

114 Although the spatial scale of GB reflects an artificial imposition onto an ecologically meaningful
115 hierarchy of scales, being neither the full range of a species nor of an individual butterfly, it reflects

116 the scale at which national policy for particular species and habitats tends to be formulated (Roy et
117 al. 2007) and at which biological recording schemes tend to be coordinated.

118 2.2. HABITAT DATA

119 Broad-leaf woodland data were obtained from the Land Cover Map 2007 (LCM2007, Morton et al.
120 2011). We chose this habitat because it is well characterised in LCM2007 and includes various
121 habitats which are prominent in UK planning and policy (e.g. ancient broad-leaf woodland, DEFRA
122 2011). The proportion of broad-leaf woodland was calculated for every 1 km grid cell in mainland GB
123 and for a 500 m radius around each UKBMS site centroid, giving a consistent scale of analysis
124 between datasets. This scale also reflects the relatively coarse resolution at which much large scale
125 habitat data is readily available. These analyses were performed in ArcGIS (v 9.3.1 © 2010 ESRI,
126 Redlands, California).

127 2.3. SCORING HABITAT ASSOCIATION FROM BIOLOGICAL RECORDING DATA

128 Analyses were performed independently. To distinguish ‘genuine’ absences for each species from a
129 1 km cell in the BNM data, as opposed to pseudoabsence generated by lack of recorders or non-
130 detection (Prendergast et al. 1993), we applied a threshold of species detection. Cells in which more
131 than five butterfly species were recorded (i.e. c. 10% of the total UK species pool, following Hickling
132 et al. (2006)) but which lacked a record of the species in question were assumed to be genuine
133 absences, whilst others were removed from all further analyses. We did not use more analytically
134 complex methods of accounting for recorder effort (e.g. Hill 2012; Isaac et al. 2014; Mason et al.
135 2015) because UK butterflies are generally well recorded, not particularly speciose, and have several
136 ubiquitous species which are well recorded across the entire of the country. Therefore, although
137 there is a latitudinal gradient in butterfly species richness in the UK, the 5 species threshold is met by
138 a relatively consistent proportion of cells per region supplementary material, Table S2). Whilst
139 butterfly species have been shown to vary in detectability (Isaac et al. 2011) there is little evidence
140 for a systematic bias whereby the detectability of individuals varies with woodland area and where
141 this relationship varies between species, which would be the only situation in which detectability
142 would automatically influence relative habitat association scores. To account for potential variation
143 in species’ habitat associations across GB, data were analysed on a regional basis, splitting the
144 dataset into 100 km by 100 km cells (from here on referred to as a 100 km region). Regions where a
145 species had less than 30 of each of presence and ‘genuine’ absence records were unlikely to provide
146 robust estimates and were excluded. We also limited analyses to species that were recorded on a
147 minimum of ten UKBMS sites.

148 General linear mixed effects models (GLMM) were used to model the relationships between habitat
149 and butterflies, using the *lme4* (Bates et al. 2013) package in R (R Core Team 2013). For the BNM
150 data, we fitted species presence/absence to proportion of broad-leaf woodland cover in the 1 km
151 grid cell, with a binomial error structure. For UKBMS data the fixed explanatory variable was
152 proportion of woodland in the 500m radius buffer whilst the response variable was total annual
153 count, adjusted for missing visits (Rothery and Roy 2001); therefore, a Poisson error structure was
154 specified. Due to the presence of zero counts for some species, we also tested analyses using zero-
155 inflated Poisson models or summing data across all years to reduce zero counts, but the species'
156 habitat association scores resulting from these models showed lower correlation with independent
157 data from expert opinion (see section 2.4). For all models, 100 km region (BNM data) or Site ID
158 (UKBMS data) was included as a random intercept, in order to account for spatial variation in the
159 mean frequency of butterflies and multiple measurements across years from the same site.
160 Preliminary analyses, comparing AIC of models with different random effect structures, also
161 supported the inclusion of a random slope whereby the relationship between proportion of
162 woodland and butterfly occurrence could vary by 100km region. For both datasets, the slope of the
163 GLMM was then designated to represent the mainland GB habitat association score, set to zero
164 where the p value was greater than 0.05.

165 To further investigate variation in habitat association by 100km region, we ran independent general
166 linear models in each region. This is more appropriate than extracting the corresponding random
167 slopes from the GLMM because of the issue of shrinkage towards the expected mean slope in
168 regions where the sample size is lower (Gelman and Hill 2007). A possible driver of spatial variation
169 in scores was investigated by performing linear regression of regional score against latitude, as
170 latitudinal gradients affect many aspects of British butterfly ecology (Oliver et al. 2009; Oliver et al.
171 2012; Oliver et al. 2014; Thomas et al. 1994; Turner et al. 1987).

172 2.4. SCORING HABITAT ASSOCIATION FROM EXPERT OPINION

173 To test the performance of the data-derived scores against established opinion, five butterfly
174 experts from research or conservation organisations (including authors TB and RF) were asked to
175 rank the species in order of woodland association, from one (strong negative association) to 50
176 (strong positive association), such that each of the 50 species could be assigned a unique rank if
177 experts deemed this suitable. Experts were requested to base rankings on where adult butterflies
178 might be expected to be encountered, rather than limiting association to breeding habitat. The
179 mean and median rankings of each species were then taken to represent average expert-derived
180 association scores for comparison with data derived scores.

2.5. DETERMINING MINIMUM SAMPLE SIZE FOR ESTIMATING HABITAT ASSOCIATION

In order to investigate the number of samples required to detect habitat associations via the data-derived methods we took random samples of presence records at a range of sample sizes, for each species. Abundance data was not re-sampled, as it showed lower correlation with expert scores (see results, section 3.2). Sample sizes analysed ranged from 100 to 1000 at intervals of 100, and from 1000 to 50000 at intervals of 500, with random sampling of occurrence records being repeated 100 times for each sample size. Each sample was then used to score habitat association using the GLMM, and the resultant scores for each sample size and species compared to expert scores. The sample size required for the ranking of the mean score from the 100 re-samplings to fall within the mean range of expert scores was then held to be the minimum sample size required for estimation of habitat association for that species (i.e. the sample size at which the ranked score is no more variable than expert scorings are from one another). We then compared these minimum sample sizes between species, and to the sample sizes typically available for species from other British taxa, applying the same selection criteria to these records as to those drawn from the BNM data (i.e. the year 1990 onwards, with 1 km precision).

3. Results

3.1. VARIATION IN EXPERT SCORES

Correlation between the habitat association scores from the two data-derived methods was significant and positive (Pearson's r , $r = 0.727$, $p < 0.001$) but with much variation in the degree of association assigned to individual species (see supplementary material, Table S1, for full table of association scores). Correlations between expert scores were always significant and strongly positive ($p < 0.001$). However, expert opinions also showed a considerable amount of variation in ranking of individual species (Figure 1). There was complete consensus in ranking only for the two highest ranked species, Purple Emperor *Apatura iris* and White Admiral *Limenitis camilla*, although other species also showed little variation in ranking - for example, Brown Hairstreak *Thecla betulae*, Silver-washed Fritillary *Argynnis paphia* and Adonis Blue *Polyommatus bellargus*.

3.2. COMPARING DATA-DERIVED SCORES WITH EXPERT SCORES

All correlation coefficients between each expert's rank score and the ranked score from occurrence data ($r = 0.646$ to 0.849) were significantly positive ($p < 0.001$) and lay within the range of correlations between experts ($r = 0.626$ to 0.909), suggesting that this method produces rankings which are no more variable from expert opinion than variation between experts. However, correlation coefficients between expert rank scores and the score from abundance data (0.554 to 0.611) were lower than all correlations between experts, suggesting that this method produced

214 rankings which varied more from expert opinion than the least concurrent pair of experts.
215 Correlations between the occurrence derived score and the mean and median expert scores ($r =$
216 0.794 , $r = 0.748$, respectively) were higher than for the abundance-derived scores ($r = 0.724$, $r =$
217 0.699 , respectively). The abundance-derived score also showed a greater number of species where
218 the data-derived ranked score fell outside the range of all expert scores (Figure 2b). These included
219 White-letter Hairstreak *Satyrrium w-album*, which was given only an intermediate ranking by the
220 abundance data but was amongst the highest ranked (i.e most strongly woodland associated) by
221 experts, and Large Heath *Coenonympha tullia*, which was also assigned an intermediate ranking by
222 the abundance data despite expert opinion giving it one of the lowest rankings. The occurrence
223 data-derived score showed fewer outliers (Figure 2a) although some species were still given rankings
224 which differed substantially from those given by experts. For example, Brown Hairstreak was ranked
225 higher by all experts than by occurrence data, whilst Marsh Fritillary *Euphydryas aurinia* was ranked
226 lower.

227 Most species showed variation in habitat association scores between 100km regions, which was in
228 many cases significantly correlated with latitude (see supplementary material, Table S1). Such
229 variation usually affected the strength of association, rather than reversing the direction of the
230 relationship. An example, for Ringlet *Aphantopus hyperantus*, is shown in Figure 3, where
231 associations were stronger in the south of GB and declined in strength with increasing latitude.

232 3.3. RE-SAMPLING TO DETERMINE MINIMUM SAMPLE SIZE FOR ESTIMATING HABITAT ASSOCIATION

233 The re-sampling of occurrence records showed that, across all species, the minimum sample size for
234 which the mean data-derived score fell within the range of expert scores had a mean of 5480
235 (standard error = ± 1750), equivalent to a mean of 223 occurrence records per 100km region.
236 However, this required minimum sample size showed considerable variation between species (see
237 supplementary material, Table S1). Species at either extreme of woodland association as
238 determined by the full-sample score and by expert opinion (i.e. with low or no significant woodland
239 association, or with high woodland association), tended to require comparatively low sample sizes
240 (100 - 1000) to come within the range of expert scores. Those species with moderate woodland
241 association scores frequently required higher sample sizes to come within the range of expert
242 scores. The mean across species was thus heavily influenced by a few species which required large
243 sample sizes, such that the mean required sample size for the ten species which showed the
244 strongest woodland associations (highest full-sample scores) was reduced to 1155 (standard error =
245 ± 815). The five most strongly woodland associated species which were suitable for analysis by re-
246 sampling (*Limenitis camilla*, *Argynnis paphia*, *Apatura iris*, *Favonius quercus*, *Leptidea sinapis*)
247 required even lower sample sizes, with a mean of 400 (standard error = ± 109).

248 **4. Discussion**

249 Our results showed that occurrence data have the potential to generate objective, quantitative
250 habitat association scores which correlate strongly with expert opinion. Scores from occurrence data
251 showed fewer deviations from expert opinion than did those from abundance data, especially for
252 specialist species (i.e. those at either extreme of the spectrum of woodland association). For
253 abundance data, the appearance of more scores which are strongly counter to expert opinion and a
254 lesser correlation with expert rankings, suggests that, invaluable though these data are for
255 monitoring population trends, they are less suitable for estimating habitat associations for certain
256 species. This may in part be an issue of statistical power, with the number of data points for
257 occurrence data (i.e. geographical locations) being orders of magnitude greater than for abundance
258 data, especially for less widespread, specialist species (e.g. Large Heath, see supplementary material,
259 Table S1). This difference in sample sizes is due to the fact that it is less intensive in terms of time
260 and effort, both in design of the monitoring scheme and in actual data collection, to acquire
261 additional occurrence data than to set up additional standardised population monitoring sites
262 (Bishop et al. 2013). There are also other issues including potential bias in the selection of locations
263 for standardised monitoring transects toward the highest quality or most accessible habitats. It thus
264 appears that in the case of assessing habitat associations, it may be better to use large quantities of
265 simple occurrence data than more detailed standardised monitoring datasets.

266 Existing, and widely used, data-derived metrics of habitat association such as IndVal (Cáceres and
267 Legendre 2009; Dufrene and Legendre 1997) compare abundance or frequency of species between
268 sites showing known differences in habitat. These rely on the location at which the organism is
269 recorded being a true reflection of the habitat with which it is associated. This is likely to be true at
270 larger spatial scales, and for sessile organisms or extensive habitats. However, many recording
271 schemes vary in the accuracy with which locations are recorded, so that the exact habitat in which
272 the species was observed is not known. In addition to this, the habitat where a species is primarily
273 found may only partly reflect the full range of resources required to complete its life cycle. In the
274 case of butterflies these include host plants, nectar plants and roosting sites (Dennis et al. 2003).
275 Our approach thus has the advantage of increasing the likelihood of capturing all essential resources
276 by testing the importance of the total proportion of a given habitat type at the landscape-scale.

277 There are still obvious limitations to this method, as not every important factor determining habitat
278 suitability is well represented by cover of a readily mapped habitat type. Such factors for butterflies
279 may include microclimates for egg laying, pupation and shelter, the presence of parasitoids or larval
280 hosts and specific resources for larval and adult feeding (Dennis et al. 2006; Dennis et al. 2003;
281 Krämer et al. 2012). Also, species are rarely restricted to only one type of habitat, and there is the

282 possibility that individuals, populations or species may adapt their habitat affinities if the primary
283 habitat is depleted or degraded (Merckx et al. 2003; Merckx and Van Dyck 2006; Proença and
284 Pereira 2013). Even where this does not occur, species may also receive benefits from habitats
285 other than the one which primarily determines their occurrence. For example, Villemey et al. (2015)
286 found that grassland butterfly richness and abundance were affected to a greater extent by local
287 woodland cover than by connectivity of the primary grassland habitat. For these reasons, the data
288 derived scores reported here should not be assumed to have captured all the information required
289 for successful species conservation. However, they should provide a robust method for assessing
290 which species are most strongly associated with a particular habitat of concern and *vice versa*, a vital
291 preliminary step in much conservation planning and policy.

292 Studies comparing or combining expert opinion with data-derived methods to optimise habitat
293 association models have shown varied results (e.g. Clevenger et al. 2002; Kuhnert et al. 2005;
294 O'Leary et al. 2009; Pearce et al. 2001; Reif et al. 2010; Seoane et al. 2005). This variation is
295 potentially driven by differences in the accuracy of expert knowledge across different locations and
296 taxa, as well as differences in the interpretation of a particular habitat type (O'Leary et al. 2009).
297 The latter probably accounts for some of the observed differences between expert opinion and
298 occurrence data in this study. For example, Purple Emperor and Brown Hairstreak are both
299 specialists of specific woodland types which comprise only a small part of the LCM2007 land cover
300 map broad-leaf woodland class - the former of extensive, mature woodlands with a tall canopy and
301 the latter of scrub and wood edge habitats, as well as hedgerows, which are not detected by
302 LCM2007.

303 Unlike many other taxa, GB butterflies are likely to be sufficiently well studied that expert opinion
304 should be well-founded, and thus a good yardstick by which to measure the performance of data-
305 derived methods. Despite this, scores varied to some extent between experts. This illustrates the
306 difficulty of using expert opinion to move beyond qualitative descriptions, even to a simple, ordinal
307 ranking for such a well-studied taxon as British butterflies. Variation amongst experts was especially
308 notable for common, widespread or mobile generalist species which received intermediate rankings
309 (Figure 1). Ranking the association of such species with a particular habitat is particularly
310 challenging, so data-derived methods may be better able to detect subtle differences in habitat
311 association between species, especially where environmental change has created differences in
312 habitat use which are not immediately apparent or where expert opinion is likely to be less well
313 informed or less up to date than for such a well monitored group as UK butterflies (Pateman et al.
314 2015; Pearce et al. 2001; Seoane et al. 2005). Experts are also likely to be most familiar with a
315 particular geographic area and base their scorings upon this knowledge. However, our results

316 (Figure 3), along with previous studies (Mayor et al. 2009; O'Leary et al. 2009; Oliver et al. 2009)
317 show the existence of spatial variation in habitat association for some species. Thus, expert opinion
318 is not necessarily transferable between geographic locations or spatial scales (Pearce et al. 2001), so
319 it may be advantageous to employ data-derived methods on data gathered over large spatial scales
320 to allow variation in habitat associations to be assessed, unless a range of experts can be canvassed
321 whose expertise cover the entire geographic area of interest. The observed spatial variation in
322 habitat association also has important implications for conservation. The Ringlet, as shown in figure
323 3, has previously been shown to exhibit shifts to core habitats under drought conditions (Sutcliffe et
324 al. 1997), so it is possible that sensitivity to drought drives the stronger affinity with woodland in
325 warmer, drier (i.e. Southern) areas of Britain, as has been demonstrated for this and other species
326 (Oliver et al. 2009; Oliver et al. 2015; Pateman et al. 2015; Suggitt et al. 2012). Such interactions
327 between habitat and climatic variables are important to consider in the light of ongoing
328 environmental change and conservation efforts to mitigate its effects (Fox et al. 2014; Oliver et al.
329 2015).

330 Examining the association with a single habitat does have the disadvantage that it is difficult to imply
331 causation – for example, a species showing a positive association with broadleaved woodland could,
332 in theory, be using a different habitat type which co-varies with woodland area (Botham et al. 2015).
333 However, whilst some significant correlations between broad habitats occurred at the regional level
334 (supplementary material, table S2), across 100km regions, there were no consistently strong
335 correlations between broadleaved woodland and any other land cover class (see supplementary
336 material, table S2), suggesting that there is no overall issue with broadleaved woodland simply being
337 a measure of some other habitat. Although this study focussed on woodland as a test case, the
338 methodology is equally applicable to any habitat (or, potentially, other environmental variables)
339 with information on spatial coverage. Analyses could thus be run for a range of land cover types to
340 find those with the highest association for each species, or by comparing scores from independent
341 models with increasing levels of habitat specificity (e.g. broad-leaf woodland, ancient broad-leaf
342 woodland, ancient oak woodlands).

343 The use of occurrence data to detect species habitat associations is likely to be most valuable for
344 other taxa for which expert opinions are likely to be more region specific, or for which there is no
345 consensus or insufficient study to form reliable expert opinions (Seoane et al. 2005). In such cases,
346 occurrence data can be relatively easily gathered from a range of sources (historical records, casual
347 species observations or national recording schemes) and so useable sample sizes may well be
348 available for a comparatively large number of species. Occurrence data also have the advantage
349 that the data collected is consistent (a species, a date and a geographical location), rather than the

350 broad range of methodologies employed in standardised monitoring schemes for different taxa,
351 such that the methods described in this study are likely to be applicable across taxa. Of course, such
352 data is only useful alongside contemporaneous environmental data, but this is becoming increasingly
353 plausible given the increasing availability of spatial environmental datasets, including digitized
354 historic mapping. Other issues associated with the use of occurrence data, particularly the need to
355 account for biases introduced by spatial and temporal accounting for recorder effort, have also
356 developed an extensive literature, with a range of methods now available (Hill 2012; Isaac et al.
357 2014; Mason et al. 2015). Such methods are likely to be a vital prerequisite in using the methods
358 described here to estimate habitat associations for poorly recorded or highly speciose groups, or
359 those with complex patterns of species richness or recorder effort.

360 The differences in the number of occurrence records required to derive habitat association scores
361 which fall within the range of those given by experts are unsurprising. It is highly likely to be easier to
362 detect stronger habitat associations at lower sample sizes. Those species requiring the largest
363 sample sizes for convergence with the expert scores were, accordingly, mostly widespread
364 generalists with moderate woodland association from the full-sample scores and expert ranking (e.g.
365 the Comma *Polygonia c-album*). Not only are larger sample sizes required to detect weak
366 relationships, but these species often showed significant spatial variation in their association scores
367 (see supplementary material, Table S1). So, whilst 5000 records might be required to ensure
368 accurate detection of subtle or cryptic habitat associations, detecting those species with strong
369 associations is likely to be possible with several hundred to 1000 records. Such species are
370 frequently those which habitat association analyses seek to identify, as being most vulnerable to
371 predicted habitat change or as potential indicators. It is also likely that robust results could be
372 obtained from lower sample sizes if there was no reason to suspect spatial variation in habitat
373 association, and therefore no reason to include a term allowing regional variation in the model.
374 However, the fact that 26% of the species analysed here showed a relationship with latitude, let
375 alone the potential for other spatial variation, suggests that accounting for spatial variation is most
376 likely necessary at all but the smallest spatial scales (Pateman et al. 2015; Pearce et al. 2001).

377 Comparing the sample sizes required to detect woodland association for butterflies with the number
378 of records for other taxa in Britain (figure 4), it is clear that butterflies are a particularly data rich
379 group (hence their use in this study as a test case). Few other taxa are as well recorded, although
380 around 30% Odonata and 10% of macro-moths meet the 5000 record threshold. For other groups,
381 although there is likely to be insufficient data to detect subtle or cryptic habitat associations,
382 comparatively large numbers of species have sufficient data to apply this method with a strong
383 probability of obtaining robust, quantitative scores for those species most reliant on a particular

384 habitat. These could then be used in a wide variety of ecological applications including the selection
385 of indicator species, the development of indices of habitat quality by weighting aggregate species'
386 population trends by degree of habitat specialisation or prediction of the extent to which each
387 species may be affected under scenarios of land-use change. Ultimately, such analyses form the
388 basis of much conservation policy at the species, habitat and ecosystem level.

389 4.1. CONCLUSIONS

390 This study has shown that analysis of recording scheme data can produce measures of habitat
391 association which support expert opinion, whilst offering several advantages over reliance on the
392 latter in terms of objectivity and the ability to detect spatial variation. The better performance of
393 readily available occurrence data over abundance data in this context confirms the value of large
394 scale volunteer recording schemes in the light of recent discussion on their comparative strengths
395 and weaknesses (Bishop et al. 2013; Dickinson et al. 2010; Tulloch et al. 2013). Although further
396 work is required to confirm the transferability of the methods detailed in this study for different
397 taxa, habitats and spatial scales, the quantitative association scores derived by the methods in this
398 study have multiple applications in conservation research.

399 **Acknowledgements**

400 We thank the experts who contributed habitat association rankings: Marc Botham, David Roy,
401 Richard Soulsby; Tom August for extraction of data for other British taxa and Damaris Zurrell for
402 constructive comments on an earlier version of the manuscript. We thank the many thousands of
403 volunteer recorders who have gathered UKBMS and BNM data. The UKBMS is operated by the
404 Centre for Ecology & Hydrology, Butterfly Conservation and the British Trust for Ornithology, and
405 funded by a multi-agency consortium including the Joint Nature Conservation Committee, Forestry
406 Commission, Natural England, the Natural Environment Research Council, Natural Resources Wales,
407 and Scottish Natural Heritage. The BNM is operated by Butterfly Conservation, with support from
408 the Centre for Ecology & Hydrology.

409 **Data Accessibility**

- 410 • The UK Land Cover Map 2007 is archived on the EIDC: Morton, R.D., Rowland, C.S., Wood, C.M.,
411 Meek, L., Marston, C.G., Smith, G.M. (2014). Land Cover Map 2007 (vector, GB) v1.2. NERC-
412 Environmental Information Data Centre doi:10.5285/2ab0b6d8-6558-46cf-9cf0-1e46b3587f13
- 413 • UKBMS and BNM data are held by Biological Records Centre on behalf of Butterfly Conservation
414 and are available on request for research purposes:
 - 415 – UKBMS data: <http://www.ukbms.org/Obtaining.aspx>

416 – BNM data: <https://data.nbn.org.uk/Datasets/GA000832>

417 **References**

- 418 Aarts, G., Fieberg, J., Brasseur, S., Matthiopoulos, J., 2013. Quantifying the effect of habitat
419 availability on species distributions. *Journal of Animal Ecology* 82, 1135-1145.
- 420 Agassiz, D.J.L., Beavan, S.D., Heckford, R.J., 2013. Checklist of the Lepidoptera of the British Isles.
421 Royal Entomological Society.
- 422 Asher, J., Warren, M., Fox, R., Harding, P., Jeffcoate, G., Jeffcoate, S., 2001. Millennium Atlas of
423 Butterflies in Britain and Ireland. Oxford University Press, Oxford
- 424 Bates, D., Maechler, M., Bolker, B., Walker, S., 2013. lme4: Linear mixed-effects models using Eigen
425 and S4. R package version 1.0-4.
- 426 Bishop, T.R., Botham, M.S., Fox, R., Leather, S.R., Chapman, D.S., Oliver, T.H., 2013. The utility of
427 distribution data in predicting phenology. *Methods in Ecology and Evolution* 4, 1024-1032.
- 428 Botham, M.S., Fernandez-Ploquin, E.C., Brereton, T., Harrower, C.A., Roy, D.B., Heard, M.S., 2015.
429 Lepidoptera communities across an agricultural gradient: how important are habitat area and
430 habitat diversity in supporting high diversity? *Journal of Insect Conservation* 19, 403-420.
- 431 Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A., Gerlach, J., Hoffmann, M., Lamoreux, J.F.,
432 Mittermeier, C.G., Pilgrim, J.D., Rodrigues, A.S., 2006. Global biodiversity conservation priorities.
433 *Science* 313, 58-61.
- 434 Cáceres, M.D., Legendre, P., 2009. Associations between species and groups of sites: indices and
435 statistical inference. *Ecology* 90, 3566-3574.
- 436 Carignan, V., Villard, M.A., 2002. Selecting indicator species to monitor ecological integrity: A review.
437 *Environmental Monitoring and Assessment* 78, 45-61.
- 438 Clevenger, A.P., Wierzchowski, J., Chruszcz, B., Gunson, K., 2002. GIS-generated, expert-based
439 models for identifying wildlife habitat linkages and planning mitigation passages. *Conservation*
440 *Biology* 16, 503-514.
- 441 DEFRA, 2011. Biodiversity 2020: A strategy for England's wildlife and ecosystem services.
- 442 Dennis, R.L., Shreeve, T.G., Van Dyck, H., 2006. Habitats and resources: the need for a resource-
443 based definition to conserve butterflies. *Biodiversity & Conservation* 15, 1943-1966.
- 444 Dennis, R.L.H., Shreeve, T.G., Van Dyck, H., 2003. Towards a functional resource-based concept for
445 habitat: a butterfly biology viewpoint. *Oikos* 102, 417-426.
- 446 Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool:
447 Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics* 41, 149-172.
- 448 Dufrene, M., Legendre, P., 1997. Species assemblages and indicator species: The need for a flexible
449 asymmetrical approach. *Ecological Monographs* 67, 345-366.

- 450 Fox, R., Oliver, T.H., Harrower, C., Parsons, M.S., Thomas, C.D., Roy, D.B., 2014. Long-term changes
451 to the frequency of occurrence of British moths are consistent with opposing and synergistic effects
452 of climate and land-use changes. *Journal of Applied Ecology* 51, 949-957.
- 453 Gelman, A., Hill, J., 2007. *Data analysis using regression and multilevel/hierarchical models*.
454 Cambridge University Press, New York ; Cambridge.
- 455 Gregory, R.D., Baillie, S.R., 1998. Large-Scale Habitat Use of Some Declining British Birds. *Journal of*
456 *Applied Ecology* 35, 785-799.
- 457 Hickling, R., Roy, D.B., Hill, J.K., Fox, R., Thomas, C.D., 2006. The distributions of a wide range of
458 taxonomic groups are expanding polewards. *Global Change Biology* 12, 450-455.
- 459 Hill, M.O., 2012. Local frequency as a key to interpreting species occurrence data when recording
460 effort is not known. *Methods in Ecology and Evolution* 3, 195-205.
- 461 Isaac, N.J.B., Cruickshanks, K.L., Weddle, A.M., Marcus Rowcliffe, J., Brereton, T.M., Dennis, R.L.H.,
462 Shuker, D.M., Thomas, C.D., 2011. Distance sampling and the challenge of monitoring butterfly
463 populations. *Methods in Ecology and Evolution* 2, 585-594.
- 464 Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen
465 science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* 5,
466 1052-1060.
- 467 Knight, J.T., Arthington, A.H., 2008. Distribution and habitat associations of the endangered Oxleyan
468 pygmy perch, *Nannoperca oxleyana* Whitley, in eastern Australia. *Aquatic Conservation: Marine and*
469 *Freshwater Ecosystems* 18, 1240-1254.
- 470 Krämer, B., Kämpf, I., Enderle, J., Poniatowski, D., Fartmann, T., 2012. Microhabitat selection in a
471 grassland butterfly: a trade-off between microclimate and food availability. *Journal of Insect*
472 *Conservation* 16, 857-865.
- 473 Kuhnert, P.M., Martin, T.G., Mengersen, K., Possingham, H.P., 2005. Assessing the impacts of grazing
474 levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environmetrics*
475 16, 717-747.
- 476 Loeb, S.C., Tainter, F.H., Cázares, E., 2000. Habitat associations of hypogeous fungi in the southern
477 Appalachians: implications for the endangered northern flying squirrel (*Glaucomys sabrinus*
478 *coloratus*). *The American Midland Naturalist* 144, 286-296.
- 479 Mason, S.C., Palmer, G., Fox, R., Gillings, S., Hill, J.K., Thomas, C.D., Oliver, T.H., 2015. Geographical
480 range margins of many taxonomic groups continue to shift polewards. *Biological Journal of the*
481 *Linnean Society* 115, 586-597.
- 482 Mayor, S.J., Schneider, D.C., Schaefer, J.A., Mahoney, S.P., 2009. Habitat Selection at Multiple Scales.
483 *Ecoscience* 16, 238-247.
- 484 Merckx, T., Dyck, H.V., Karlsson, B., Leimar, O., 2003. The evolution of movements and behaviour at
485 boundaries in different landscapes: a common arena experiment with butterflies. *Proceedings of the*
486 *Royal Society of London B: Biological Sciences* 270, 1815-1821.
- 487 Merckx, T., Van Dyck, H., 2006. Landscape structure and phenotypic plasticity in flight morphology in
488 the butterfly *Pararge aegeria*. *Oikos* 113, 226-232.

- 489 Morton, D., Rowland, C., Wood, C., Meek, L., Marston, C., Smith, G., Simpson, I.C., 2011. Final report
490 for LCM2007 - the new UK land cover map. , p. 112pp. NERC/Centre for Ecology and Hydrology.
- 491 O'Leary, R.A., Choy, S.L., Murray, J.V., Kynn, M., Denham, R., Martin, T.G., Mengersen, K., 2009.
492 Comparison of three expert elicitation methods for logistic regression on predicting the presence of
493 the threatened brush-tailed rock-wallaby *Petrogale penicillata*. *Environmetrics* 20, 379-398.
- 494 Oliver, T., Hill, J.K., Thomas, C.D., Brereton, T., Roy, D.B., 2009. Changes in habitat specificity of
495 species at their climatic range boundaries. *Ecology Letters* 12, 1091-1102.
- 496 Oliver, T.H., Marshall, H.H., Morecroft, M.D., Brereton, T., Prudhomme, C., Huntingford, C., 2015.
497 Interacting effects of climate change and habitat fragmentation on drought-sensitive butterflies.
498 *Nature Clim. Change* 5, 941-945.
- 499 Oliver, T.H., Roy, D.B., Brereton, T., Thomas, J.A., 2012. Reduced variability in range-edge butterfly
500 populations over three decades of climate warming. *Global Change Biology* 18, 1531-1539.
- 501 Oliver, T.H., Stefanescu, C., Paramo, F., Brereton, T., Roy, D.B., 2014. Latitudinal gradients in
502 butterfly population variability are influenced by landscape heterogeneity. *Ecography* 37, 863-871.
- 503 Pateman, R.M., Hill, J.K., Roy, D.B., Fox, R., Thomas, C.D., 2012. Temperature-Dependent Alterations
504 in Host Use Drive Rapid Range Expansion in a Butterfly. *Science* 336, 1028-1030.
- 505 Pateman, R.M., Thomas, C.D., Hayward, S.A.L., Hill, J.K., 2015. Macro- And Micro-Climatic
506 Interactions Can Drive Variation in Species' Habitat Associations. *Global Change Biology*, n/a-n/a.
- 507 Pearce, J.L., Cherry, K., Drielsma, M., Ferrier, S., Whish, G., 2001. Incorporating expert opinion and
508 fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied*
509 *Ecology* 38, 412-424.
- 510 Pollard, E., Yates, T.J., 1993. *Monitoring Butterflies for Ecology and Conservation*. Chapman & Hall,
511 London.
- 512 Prendergast, J.R., Wood, S.N., Lawton, J.H., Eversham, B.C., 1993. Correcting for Variation in
513 Recording Effort in Analyses of Diversity Hotspots. *Biodiversity Letters* 1, 39-53.
- 514 Proença, V., Pereira, H.M., 2013. Species–area models to assess biodiversity change in multi-habitat
515 landscapes: The importance of species habitat affinity. *Basic and Applied Ecology* 14, 102-114.
- 516 R Core Team, 2013. *R: A language and environment for statistical computing*. R Foundation for
517 Statistical Computing, Vienna, Austria.
- 518 Reif, J., Jiguet, F., Šťastný, K., 2010. Habitat specialization of birds in the Czech Republic: comparison
519 of objective measures with expert opinion. *Bird Study* 57, 197-212.
- 520 Rothery, P., Roy, D.B., 2001. Application of generalized additive models to butterfly transect count
521 data. *Journal of Applied Statistics* 28, 897-909.
- 522 Rouquette, J.R., Thompson, D.J., 2005. Habitat associations of the endangered damselfly, *Coenagrion*
523 *mercuriale*, in a water meadow ditch system in southern England. *Biological Conservation* 123, 225-
524 235.

525 Roy, D.B., Rothery, P., Brereton, T., 2007. Reduced-effort schemes for monitoring butterfly
526 populations. *Journal of Applied Ecology* 44, 993-1000.

527 Seoane, J., Bustamante, J., DÍAz-Delgado, R., 2005. Effect of Expert Opinion on the Predictive Ability
528 of Environmental Models of Bird Distribution

529 Efecto de la Opinión de Experto en la Capacidad Predictiva de Modelos de Distribución de Aves
530 usando Predictores Ambientales. *Conservation Biology* 19, 512-522.

531 Suggitt, A.J., Stefanescu, C., Páramo, F., Oliver, T., Anderson, B.J., Hill, J.K., Roy, D.B., Brereton, T.,
532 Thomas, C.D., 2012. Habitat associations of species show consistent but weak responses to climate.
533 *Biology Letters*.

534 Sutcliffe, O.L., Thomas, C.D., Yates, T.J., Greatorex-Davies, J.N., 1997. Correlated extinctions,
535 colonizations and population fluctuations in a highly connected ringlet butterfly metapopulation.
536 *Oecologia* 109, 235-241.

537 Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus,
538 B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S.,
539 Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Townsend Peterson, A., Phillips, O.L., Williams, S.E.,
540 2004. Extinction risk from climate change. *Nature* 427, 145-148.

541 Thomas, J., 2005. Monitoring change in the abundance and distribution of insects using butterflies
542 and other indicator groups. *Philosophical Transactions of the Royal Society B: Biological Sciences*
543 360, 339-357.

544 Thomas, J.A., Moss, D., Pollard, E., 1994. Increased fluctuations of butterfly populations towards the
545 northern edges of species' ranges. *Ecography* 17, 215-220.

546 Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013. Realising the full
547 potential of citizen science monitoring programs. *Biological Conservation* 165, 128-138.

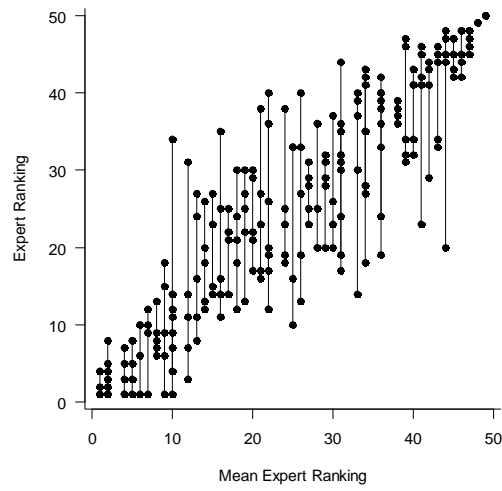
548 Turner, J.R.G., Gatehouse, C.M., Corey, C.A., 1987. Does Solar Energy Control Organic Diversity?
549 Butterflies, Moths and the British Climate. *Oikos* 48, 195-205.

550 Villemey, A., van Halder, I., Ouin, A., Barbaro, L., Chenot, J., Tessier, P., Calatayud, F., Martin, H.,
551 Roche, P., Archaux, F., 2015. Mosaic of grasslands and woodlands is more effective than habitat
552 connectivity to conserve butterflies in French farmland. *Biological Conservation* 191, 206-215.

553 Warren, M.S., Hill, J.K., Thomas, J.A., Asher, J., Fox, R., Huntley, B., Roy, D.B., Telfer, M.G., Jeffcoate,
554 S., Harding, P., Jeffcoate, G., Willis, S.G., Greatorex-Davies, J.N., Moss, D., Thomas, C.D., 2001. Rapid
555 responses of British butterflies to opposing forces of climate and habitat change. *Nature* 414, 65-69.

556 Yapp, R.H., 1922. The Concept of Habitat. *Journal of Ecology* 10, 1-17.

557



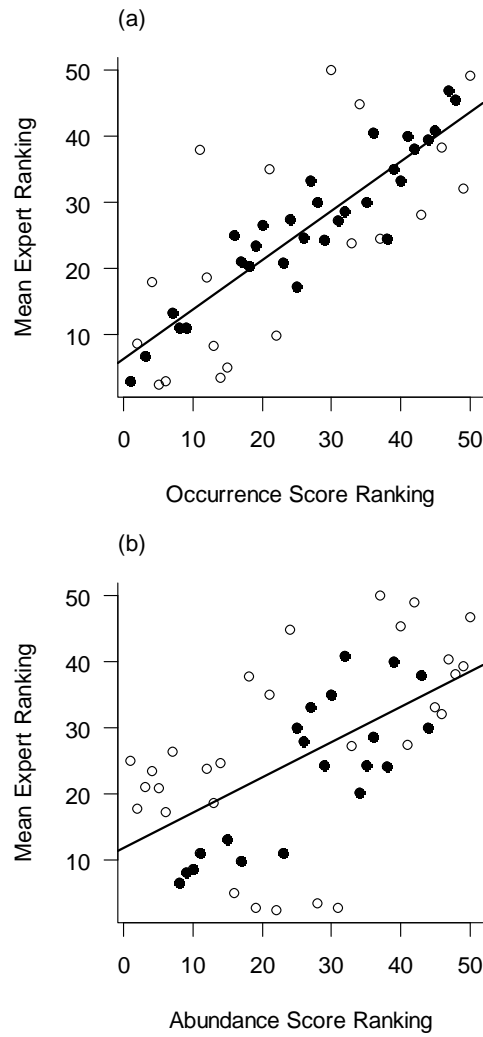
559

560 **Fig.1** Plot of variation in expert rankings of butterfly woodland association scores. The mean expert

561 ranking is plotted on the x axis, with the associated rankings given by each expert plotted on the y

562 axis. Vertical black lines indicate the range of rankings across all five experts for each species.

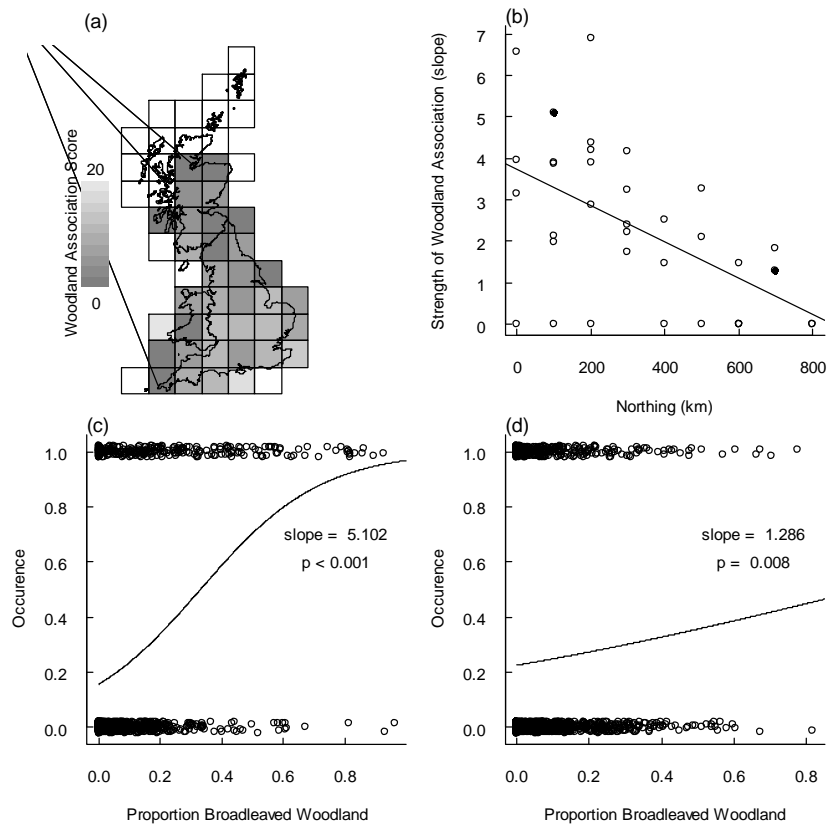
563



565

566 **Fig. 2** Plots of a) Rank score from occurrence data and b) Rank score from abundance data against
567 mean expert ranking. Open circles are species for which the ranking of the data-derived score did
568 not lie within the range of the rankings assigned by experts.

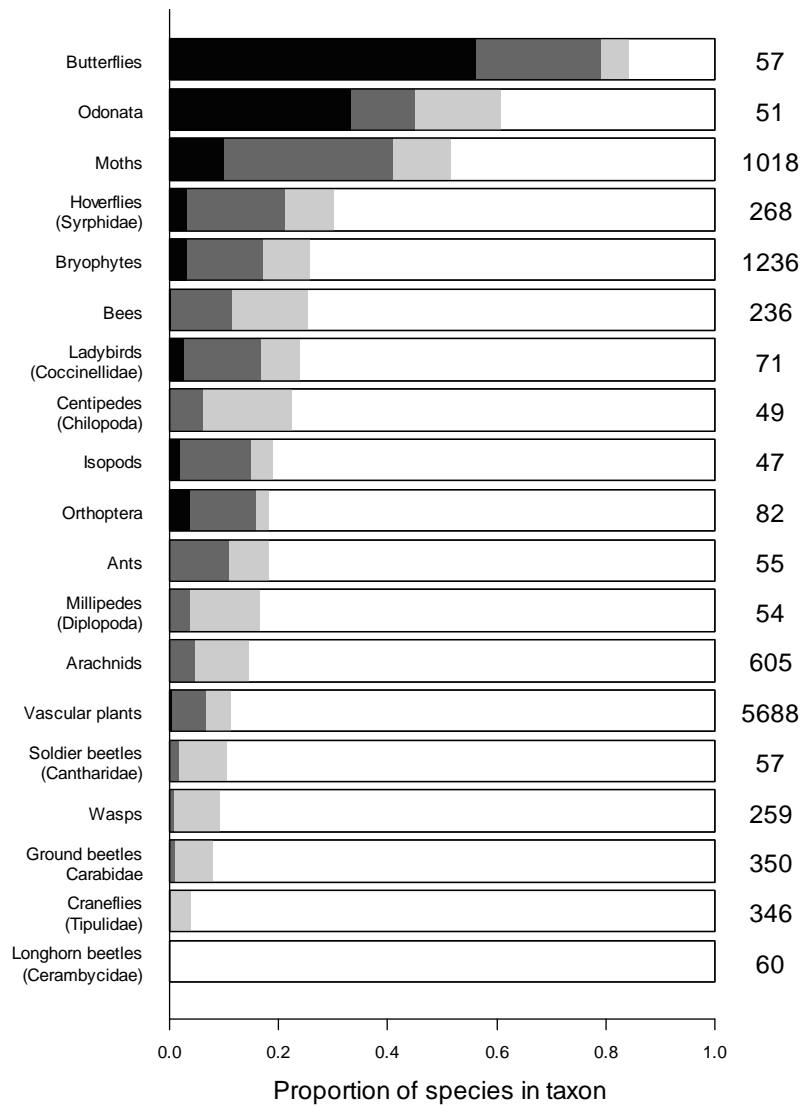
569



570

571 **Fig.3** Regional variation in the association of the Ringlet butterfly *Aphantopus hyperantus* with
 572 broad-leaf woodland across Great Britain (GB), as detected from occurrence data. (a) Map of GB
 573 showing relative strength of association in 100 km regions. Unshaded regions had insufficient data
 574 for analysis. (b) Plot of association scores against latitude, measured as distance north from grid
 575 origin (Northing). Filled points indicate example regions where the relationship is shown in panel (c),
 576 an example of a strong, positive relationship with broad-leaf woodland and panel (d), an example of
 577 a weaker relationship with broad-leaf woodland.

578



579

580

Fig 4. Comparisons of proportions of species in different GB taxa which meet various thresholds in

581

number of unique occurrence records with 1 km or better precision. Numbers to the right of bars

582

indicate total number of species within each taxon. Sections of bars are shaded by number of

583

species meeting thresholds: black sections = 5000 records, dark grey sections = 1000 records, light

584

grey sections = 500 records.

585 **Supplementary material**

586 Additional supplementary material may be found in the online version of this article:

587 **Table S1** Results of scoring species association with broad-leaf woodland for 50 butterfly species in
588 mainland GB, from abundance data, occurrence data and expert opinion.

589 **Table S2** Pearson's correlation coefficients from correlations between broadleaved woodland and
590 other land cover classes from the UK land cover map 2007, by 100 km x 100 km region.