

# Characterisation of false-positive observations in botanical surveys (#10473)

1

First submission

Please read the **Important notes** below, the **Review guidance** on page 2 and our **Standout reviewing tips** on page 3. When ready [submit online](#). The manuscript starts on page 4.

## Important notes

### Editor and deadline

Richard Cowling / 19 Mar 2017

### Files

6 Figure file(s)

2 Table file(s)

Please visit the overview page to [download and review](#) the files not included in this review PDF.

### Declarations

No notable declarations are present



Please read in full before you begin

## How to review

When ready [submit your review online](#). The review form is divided into 5 sections. Please consider these when composing your review:

### 1. BASIC REPORTING

### 2. EXPERIMENTAL DESIGN

### 3. VALIDITY OF THE FINDINGS






4. General comments

5. Confidential notes to the editor





 You can also annotate this PDF and upload it as part of your review

To finish, enter your editorial recommendation (accept, revise or reject) and submit.





### BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [PeerJ standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [PeerJ policy](#)).

### EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

### VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  Conclusions are well stated, linked to original research question & limited to supporting results.
-  Speculation is welcome, but should be identified as such.
-  Data is robust, statistically sound, & controlled.

The above is the editorial criteria summary. To view in full visit <https://peerj.com/about/editorial-criteria/>

# 7 Standout reviewing tips

3



The best reviewers use these techniques

## Tip

## Example

**Support criticisms with evidence from the text or from other sources**

*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*

**Give specific suggestions on how to improve the manuscript**

*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*

**Comment on language and grammar issues**

*The English language should be improved to ensure that your international audience can clearly understand your text. I suggest that you have a native English speaking colleague review your manuscript. Some examples where the language could be improved include lines 23, 77, 121, 128 - the current phrasing makes comprehension difficult.*

**Organize by importance of the issues, and number your points**

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

**Give specific suggestions on how to improve the manuscript**

*Line 56: Note that experimental data on sprawling animals needs to be updated. Line 66: Please consider exchanging "modern" with "cursorial".*

**Please provide constructive criticism, and avoid personal opinions**

*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*

**Comment on strengths (as well as weaknesses) of the manuscript**

*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

# Characterisation of false-positive observations in botanical surveys

Quentin J Groom <sup>Corresp., 1</sup>, Sarah J. Whild <sup>2</sup>

<sup>1</sup> Botanic Garden Meise, Meise, Belgium

<sup>2</sup> School of Science and the Environment, The Manchester Metropolitan University, Manchester, United Kingdom

Corresponding Author: Quentin J Groom  
Email address: qgroom@reticule.co.uk

Errors in botanical surveying are a common problem. The presence of a species is easily overlooked, leading to false-absences; while misidentifications and other mistakes lead to false-positive observations. While it is common knowledge that these errors occur, there are few data that can be used to quantify and describe these errors. Here we characterise false-positive errors for a controlled set of surveys conducted as part of a field identification test of botanical skill. Surveys were conducted at sites with a verified list of vascular plant species. The candidates were asked to list all the species they could identify in a defined botanically rich area. They were told beforehand that their final score would be the sum of the correct species they listed, but false-positive errors counted against their overall grade. The number of errors varied considerably between people, some people create a high proportion of false-positive errors, but these are scattered across all skill levels. Therefore, a person's ability to correctly identify a large number of species is not a safeguard against the generation of false-positive errors. There was no phylogenetic pattern to falsely observed species, however, rare species are more likely to be false-positive as are species from species rich genera. Raising the threshold for the acceptance of an observation reduced false-positive observations dramatically, but at the expense of more false negative errors. False-positive errors are higher in field surveying of plants than many people may appreciate. Greater stringency is required before accepting species as present at a site, particularly for rare species. Combining multiple surveys resolves the problem, but requires a considerable increase in effort to achieve the same sensitivity as a single survey. Therefore, other methods should be used to raise the threshold for the acceptance of a species. For example, digital data input systems that can verify, feedback and inform the user are likely to reduce false-positive errors significantly.

1 Characterisation of false-positive observations in botanical  
2 surveys

3

4 Quentin J. Groom

5 Botanic Garden Meise, Bouchout Domain, Nieuwelaan 38, 1860 Meise, Belgium

6 Email: [quentin.groom@plantentuinmeise.be](mailto:quentin.groom@plantentuinmeise.be)

7 Tel: +32 2 260 09 20 ext. 364

8

9 Sarah J. Whild

10 School of Science and the Environment, Manchester Metropolitan University, The Gateway,  
11 Shrewsbury, SY1 1NB, United Kingdom.


12

### 13 Abstract

14 Errors in botanical surveying are a common problem. The presence of a species is easily  
15 overlooked, leading to false-absences, whilst misidentifications and other mistakes lead to  
16 false-positive observations. While it is common knowledge that these errors occur, there are  
17 few data that can be used to quantify and describe these errors. Here we characterise false-  
18 positive errors for a controlled set of surveys conducted as part of a field identification test of  
19 botanical skill. Surveys were conducted at sites with a verified list of vascular plant species.  
20 The **candidates** were asked to list all the species they could identify in a defined botanically  
21 rich area. They were told beforehand that their final score would be the sum of the correct  
22 species they listed, but false-positive errors counted against their overall grade. The number  
23 of errors varied considerably between **people**, some people create a high proportion of  
24 false-positive errors, but these are scattered across all skill levels. Therefore, a **person's**  
25 ability to correctly identify a large number of species is not a safeguard against the  
26 generation of false-positive errors. There was no phylogenetic pattern to falsely observed  
27 species, however, rare species are more likely to be false-positive as are species from  
28 species rich genera. Raising the threshold for the acceptance of an observation reduced  
29 false-positive observations dramatically, but at the expense of more false negative errors.  
30 False-positive errors are higher in field surveying of plants than many people may  
31 appreciate. Greater stringency is required before accepting species as present at a site,  
32 particularly for rare species. Combining multiple surveys resolves the problem, but requires  
33 a considerable increase in effort to achieve the same sensitivity as a single survey.  
34 Therefore, other methods should be used to raise the threshold for the acceptance of a  
35 species. For example, digital data input systems that can verify, feedback and inform the  
36 user are likely to reduce false-positive errors significantly.

## 37 Introduction





38 Errors in science are inevitable. Sometimes they are the result of random chance but more often  
39 than not, they are the result of human fallibility. Errors are particularly common in observations  
40 of biodiversity, where organisms can be either inconspicuous, hard to identify or hidden.  
41 Furthermore, the process of observation can be disrupted by external influences and observer  
42 biases (Simons et al., 2007; Willson et al., 2011). Animals are often intentionally secretive, but  
43 even sedentary organisms, such as plants, are difficult to observe owing to their similarity to  
44 each other. These sorts of errors lead to false-negative errors. False-negative errors are  
45 expected in plant surveys due to the variability in the detectability of different species in different  
46 seasons and habitats (Rich & Woodruff, 1992; Chen et al., 2009, 2013). False-positive errors  
47 however, are those arising from observing something that was not there. These errors have  
48 been given many names, including detection errors, type 1 errors, errors of commission and  
49 misclassifications. Here we have chosen to use the terms false-positive and false-negative for  
50 the sake of readability. False-positive errors occur for several reasons; people can either  
51 misidentify an organism; wrongly report the date or location or incorrectly transcribe otherwise  
52 correctly reported data.

53  There are few studies on errors in botanical recording and few survey schemes have specific  
54 quality assurance mechanisms, such as suggested by Scott & Hallam (2003). In contrast, much  
55 more attention has been paid to observation errors of animals where progress has been made  
56 in the methods for observation and analysis (Elphick, 2008). Much of expert plant identification  
57 in the field is done using gestalt perception, rather than formal identification of characters (Ellis,  
58 2011). However, human senses and reasoning, though remarkable, are prone to various sorts  
59 of error including apophenia, generalizations and confirmation bias.

60 Users of botanical records expend considerable effort in “cleaning” data (Chapman, 2005).

61 Cleaning entails using experience to verify and reformat the results of biological surveys,

62 however, this is an inefficient process as errors are much better resolved close to their sources.  
63 Furthermore, data “cleaning” is also fallible, and elimination of errors early in the workflow is  
64 likely to be quicker and less costly than when time and distance is put between the observation  
65 event and the person analysing it. Statistical methods can also be used to account for observer  
66 errors and bias (Miller et al., 2011; Bird et al., 2014; Dorazio, 2014), but while these approaches  
67 are useful, the first line of defence should be minimizing errors in field surveying.


68 In biological surveying, false-positive observations are arguably more costly than false-  
69 negatives. If false-negatives are suspected, additional surveying can help to resolve them,    
70 false-negatives are more likely to occur for rare species thus the more surveying conducted the  
71 more confident   me that the species is either truly absent or at least extremely rare  
72 within  survey area. Indeed, as an absence can never be proven, evidence for extreme  
73 rarity is the best we can hope for. In the case of threatened species, false-negatives could lead  
74 to inappropriate actions in planning decisions or site management, but false-positives give the  
75 impression that a species is more common than it is and may lead to its conservation status not  
76 being recognised.



77 In contrast to false-negatives, false-positive errors are difficult to refute and can pollute datasets  
78 indefinitely. One only has to think of the time and effort wasted on extreme cases of false-  
79 positive errors, such as observations of plesiosaurs in Loch Ness and hominids in the Rocky  
80 Mountains, but there are many other examples (Sabbagh, 2001; McKelvey et al., 2008).

81 In this study we use a quite unique set of plant surveying data where the same sites have been  
82 surveyed repeatedly by many independent observers. These data are derived from tests for  
83 Field Identification Skills Certificates (FISCs) that have been running for eight years under the  
84 aegis of the Botanical Society of Britain and Ireland. These certificates are intended to give the  
85 participants and potentially their future or current employers a guide to their skill at vascular



86 plant identification in the field. These day-long tests include two laboratory-based tests and an  
87 afternoon field test and it is from this field test that the data in this paper are derived.

88 In classical signal detection theory, where the signal has to be separated from the noise, we can  
89 reduce the number of false-positive errors by increasing the detection threshold for accepting  
90 the signal. However, this is at the expense of an increase in false-negative errors (Wolf et al.,  
91 2013). We can examine this trade-off by changing the acceptance criteria for a species to be  
92 present. One simple method to increase the detection threshold is to combine the results of two  
93 or more observers, only accepting observations if multiple observers agree. So called  double-  
94 observer methods are frequently used in animal surveys, particularly for avian point counts  
95 (Simons et al., 2007; Conn et al., 2013).

96 Double-observer methods can reduce the false-positive observations because false-positives  
97 are rare. If false-positives always account for  all proportion of the total number  observations  
98 and an observer picks their false-positive observations randomly from a fairly large pool of  
99 species names, then the number of false-positive observations of any one species should  
100 always be small and the chances of two observers picking the same false-positive species is  
101 extremely small.

102 However, there are two potential problems with this approach. Firstly, all the species that are  
103 actually present but only observed by one observer become false-negative observations.  
104 Secondly, the assumption that observers are unbiased in their creation of false-positive  
105 observations may not be true. For example, surveyors may pick false-positive observations from  
106 species closely related to those actually at the site or they may pick them from common plants  
107 that they assume to be present.

108 In this paper we examine the characteristics of false-positive observations from the perspective  
109 of plant detectability, relatedness and their familiarity to observers. We examine whether  
110 changing the threshold for a true positive observation improves the accuracy of surveys and we

111 discuss what other strategies could be used to reduce errors in botanical surveying. The  
112 intention is that the results can be used to design better plant survey methods that will lead to a  
113 reduction in the number of false-positive observations.

## 114 **Materials & Methods**

### 115 **Sites description**

116 This analysis is based on the field test data collected from 238 surveys from the FISCs  
117 conducted in Shropshire, UK. From 2007 to 2014 six sites have been used, Sweeney Fen;  
118 Ballstone Quarry; Windmill Hill; Aston Locks off-line reserve; The Old River Bed, Shrewsbury  
119 and Blakeway Hollow. A summary of the sites and the surveys conducted on them are  
120 presented in table 1. The sites were chosen to fit the following criteria, which are consistent with  
121 the FISC protocols. They are around 2–3 hectares of accessible habitat, which are relatively  
122 safe in health and safety terms. They were selected for their habitat, such as unimproved  
123 grassland with some scrub areas, or short fen, or not too wet sedge swamp, or broad-leaved  
124 woodland. The area to be surveyed was made very clear to the participants – if a smaller area  
125 was used within a larger reserve it was fenced or taped off clearly. It was made clear to  
126 participants whether or not hedges were to be surveyed.

127 The site selection criteria were as follows.

- 128 • A relatively small more or less homogeneous site with fairly distinct boundaries.
- 129 • Small enough to survey thoroughly within two hours.
- 130 • Large enough for individual surveys to be carried out and for invigilation to be effective.
- 131 • Possessing a reasonably complex vegetation with a good range of grasses, sedges  
132 and/or rushes, giving a total of around 100 vascular plant species to record.

133 Windmill Hill and Ballstone Quarry are grassland sites on Silurian Limestone. Blakeway Hollow  
134 is also on Silurian Limestone but is a sunken trackway with grass verges and also dense

135 species-rich hedges. Sweeney Fen is a mixture of neutral grassland and calcareous fen over  
136 Carboniferous Limestone. The Old River Bed is an old meander filled with sedge swamp on  
137 Quaternary deposits. Aston Locks off-line reserve is adjacent to the Montgomery Canal and has  
138 areas of open water, neutral grassland, tall herbs and some sedge swamp.

139 All sites were relatively easy to access over stiles or through gates. The Old River Bed is  
140 arguably the most challenging site as it can be wet, and Windmill Hill is on a steep slope but the  
141 sites were chosen to be as accessible as possible.

142 Demographic data on surveyors was not collected, but anecdotally the age range is between 25  
143 and 60 with a median in the 30s. The gender balance is roughly equal. The main motivation for  
144 participants for taking a FISC has been career enhancement, with ecological consultants  
145 forming the bulk of participants. Occasionally, surveyors repeated the FISC, after gaining some  
146 experience in the field, however they were never tested on the same site twice.

147

148 The FISC field tests were evaluated by the number of correct species recorded in the field, and  
149 a score based on false-positive errors. Candidates were clearly informed before the test that  
150 they would lose marks for false-positive observations giving them a clear disincentive to make  
151 mistakes.

152 The species lists produced in the field by FISC participants were of two sorts – participants were  
153 offered a choice of using a pre-printed survey card with abbreviated scientific names of the  
154 species on the card were restricted to Shropshire and did not include very rare species. The  
155 other option was recording free hand, using unambiguous common names or scientific names.  
156 Candidates were permitted to use whatever literature they wished for identification purposes  
157 and they were not penalised for using standard vernacular names or synonyms.

158 FISC participants recorded against a 'gold standard' surveyor – a volunteer at skill level 5 (level  
159 5 is a professional level plant recorder who is competent to record in most habitats and areas in  
160 the UK) who recorded under the same conditions for the same length of time and scores for  
161 participants were calculated as a percentage of the 'gold standard' surveyor's total.

## 162 Data

163 The digitisation of the field data was carried out by a small team with instructions to pick out four  
164 categories of botanical records – correct species; unreasonable or known incorrect; 'mythical'  
165 species and 'cautious' records, with just a generic name. Data were simplified to species with  
166 taxonomy following Stace (2010), except for the phylogenetic signal tests where the taxonomy  
167 was aligned with the Daphne phylogeny (Durka & Michalski, 2012). The number of species for  
168 each genus was taken from Stace (2010), where species for each genus in the UK and Ireland  
169 are numbered.

170 The data used in this paper has been openly deposited in the Zenodo repository under DOI  
171 10.5281/zenodo.46662.

## 172 Bootstrapping scripts

173 To measure the average numbers of false-positive and false-negative observations when  
174 surveys were combined a bootstrapping approach was used. Depending on the number of  
175 surveys to be tested, a random selection of surveys was selected from the pool of all surveys  
176 conducted at each site. The number of false-positives and true positives were calculated from  
177 this selection and average values were calculated from 10,000 randomly chosen selections.  
178 This script was written in Perl (ActivePerl, version 5.16.3.1603) and is available in a public  
179 Github repository (<https://github.com/qgroom/grouped-surveys>).

## 180 Statistics

181 Statistics were performed in R version 3.1.0. Owing to the differences in species composition,  
182 species abundance and habit at each site, all analyses were conducted on each site separately  
183 and treated as individual replicates. Generalized linear models using the total number of false-  
184 positive observations resulted in overdispersed models, because so many of the species have  
185 either zero or one false-positive observation. The solution was to model the false-positives as a  
186 binary response variable, with a species either being a true-positive (zero) or a false-positive  
187 (one). The proportion of false-positive observations was modelled using a generalized linear  
188 model using a logit link function. The mean interpolated 4 km<sup>2</sup> occupancy probability from  
189 southern England between 1995 and 2011 was used as a measure of the regional frequency of  
190 the species (Groom, 2013).

191 There are several tests for phylogenetic signal strength, particularly for continuous traits.  
192 However, as the majority of species had either one or zero false-positive observations we again  
193 chose to treat false-positives as a binary trait. To test for a phylogenetic component to the false  
194 detection of species the D statistic was calculated following Fritz & Purvis (2010). The D statistic  
195 is a measure of the phylogenetic signal strength of binary traits. The D statistic was calculated  
196 using the Caper package version 0.5.2 (Orme, 2012). The phylogeny used for the calculation  
197 was the Daphne phylogeny of Northern European plants (Durka & Michalski, 2012). To  
198 calculate the D statistic for generic observations the Daphne tree was pruned using the Phytools  
199 package (Revell, 2012).

200 Mean specificity and sensitivity were calculated from the mean bootstrapped values for true  
201 positive, false-negative and false-positive observations. True negative observations were  
202 calculated from the cumulative number of false-positive species recorded at the site, minus the  
203 number of false-positive observations. Using this value for the number of true negative  
204 observations is somewhat arbitrary, as any number of species are truly absent from the site.  
205 However, using other values only affects the absolute value of the specificity, not their relative

206 value. Sensitivity is calculated from the number of true positive observations, divided by the sum  
207 of the true positive observations and the false-negatives. Specificity is calculated from the  
208 number of true negatives, divided by the sum of the true negatives and false-positives.

209 All confidence intervals quoted in the text are calculated using the t distribution.

## 210 Results

211 The number of false-positive errors is not strongly related to the number of correct observations  
212 a surveyor makes (Fig. 1). Across all sites the number of false-positive errors is not significantly  
213 correlated with the number of correct observations ( $R^2 = 0.14$  [95% C.I.  $-0.11 - 0.40$ ,  $n=6$ ]).

214 Experts, who can identify many species, are not necessarily likely to create fewer false-positive  
215 errors. Some surveyors are cautious while others are reckless.

216 Fig. 2 demonstrates that false-positives have a much lower detection probability than true  
217 positives. Based on the raw data, surveyors generate an average of 3.4% false-positive  
218 observations per survey. However, we have to accept that we could have made errors in our  
219 assessment of what species are at each site, which could either increase or decrease the  
220 number of apparent false-positive errors for the candidates.

221 The detection probabilities of species accepted to be at the site vary widely, from close to one to  
222 close to zero (Fig. 2), though about 75% of species have a detection probability less than 0.5.

223 The false detection probability is always less than 0.5 and the majority of false detection  
224 probabilities are less than 0.05. Again, we should consider that the reference survey may have  
225 contained some false-negative errors, resulting in apparent false-positive errors for the

226 candidates. In such cases it is likely that several of the candidates will have observed these  
227 species and this may explain the right-hand tail of higher false detection probabilities seen in  
228 Fig. 2.

229 To investigate whether surveyors were more likely to make mistakes if there were many species  
230 to choose from in a genus we correlated the total number of false-positive observations for  
231 members of a genus and the number of species in a genus in the UK. At all six sites this gave a  
232 significant positive correlation ( $R^2 = 0.34$  [95% C.I. 0.59 – 0.10, n=6]).

### 233 **Is there a phylogenetic component to the false detection of species?**

234 Values of D usually vary between 0 and 1 and are inversely proportional to the degree of  
235 phylogenetic clustering. Values less than zero are expected if phylogenetically closely related  
236 species are more often identified as false-positive errors. Values greater than zero occur where  
237 there is overdispersion, such that closely related species show opposite results. This would be  
238 the case where observers create errors that are closely related to the correct species. The D  
239 statistic is compared to the results from two evolutionary models, random and Brownian. Values  
240 close to 1 indicate that the random model is most appropriate and there is no phylogenetic  
241 relationship. The D statistic for false-positive observations at all six sites was close to one,  
242 indicating that there is no phylogenetic signal in the false detection of species (mean 1.04 [95%  
243 C.I. 0.98-1.11], n=6) (Table 2).

### 244 **Does familiarity with a species influence the false detection of species?**

245 By comparing the relationship between the numbers of species chosen as false-positives with  
246 the average 4 km<sup>2</sup> occupancy for southern England, we can compare how commonness and  
247 rarity relate to the likelihood of mis-observation. For each site the proportion of false-positive  
248 species is more for rarer species (Fig. 3). From the intercept of these models we can also  
249 estimate the probability of observing species that do not occur in southern England, which is  
250 0.237 [95% C.I. 0.184 – 0.291, n=6]; though this seems a large proportion, this value is  
251 cumulative for all surveys conducted at the site.

### 252 **Inseparable taxa and fictional taxa**

253 If surveyors were not able to identify a plant to species they could alternatively report a genus.  
254 Unlike errors these are conscious decisions to record a taxon at a less resolved taxon rank. The  
255 ten most recorded genera in this category were *Rosa*, *Salix*, *Euphrasia*, *Viola*, *Carex*,  
256 *Equisetum*, *Hypericum*, *Quercus*, *Epilobium* and *Myosotis*; listed in decreasing order of their  
257 number of observations. For all six sites there was a weak, but significant, positive correlation  
258 between the number of cautious observations and the number of species in a genus ( $R^2 = 0.19$   
259 [95% C.I. 0.27 – 0.10, n=6]). We did wonder whether surveyors who recorded more cautious  
260 taxa would record fewer false-positives. However, there is no evidence for this in these data.  
261 The number of cautious taxa recorded by a surveyor is not significantly correlated with their  
262 number of false-positive observations ( $R^2 = -0.16$  [95% C.I. -0.19 – 0.07, n=6]).

263 No significant phylogenetic pattern was seen in the recording of genera. The D statistic was  
264 calculated for all sites separately. The detailed results are not shown but the average D across  
265 sites was 0.85 [95% C.I. 0.62–1.07] n = 6.

266 In addition to cautious identifications there were non-existent taxa. These were a mixture of  
267 imaginary Latin and English names. None were completely imaginary and most were  
268 combinations of words that exist in other names, but not together. For example, there were  
269 incorrect Latin combinations such as *Calystegia arvense*, *Carex articulatus*, *Geranium palustre*,  
270 *Plantago ovalifolium* and *Polygonum vulgare*. Imagined vernacular names included black  
271 alkanet, burweed, separated rush and sheep's-beard. Names such as burweed do exist for  
272 species that do not occur in the British Isles, but it is assumed that the species intended by the  
273 observer was one that occurs in the British Isles with burred fruits, such as *Galium aparine*,  
274 *Geum urbanum* and *Torilis japonica*.

## 275 An evaluation of combining the results of the multiple surveys

276 Increasing the threshold for acceptance of an observation dramatically reduced the number of  
277 false-positive observations (Fig. 4). Even when only two surveys are combined the average



278 number of false-positive observations is 8% of the number in a single survey and false-positive  
279 observations are practically eliminated if five surveys are combined.

280 Two useful metrics for test performance are the specificity and sensitivity. The specificity, in the  
281 context of botanical surveying, is the probability that a species is not seen given the plant is not  
282 present, it decreases with increasing numbers of false-positive observations. The sensitivity is  
283 the probability that the plant is observed given that the plant is present, it decreases with  
284 increasing numbers of false-negative errors. Figures 5 and 6 show how the average specificity  
285 and sensitivity of plant observation change with different numbers of repeat surveys, using an  
286 acceptance threshold of presence in two surveys. With just two surveys the specificity of  
287 surveying is close to its maximum value (Fig. 5). However, the increase in specificity has been  
288 at the expense of sensitivity, as combining two surveys creates many false-negatives (Fig. 6).  
289 Yet, with additional surveys the sensitivity rises quickly and exceeds the sensitivity of a single  
290 survey, whereas the specificity declines only slowly. With three surveys the sensitivity is  
291 approximately the same as a single survey, but the specificity has improved considerably.

292 A common strategy for the analysis of botanical records is to combine the results of multiple  
293 surveys, so that a species only has to be observed once to be included. The sensitivity of such  
294 a strategy increases with increasing numbers of surveys (Fig. 6). However, owing to the  
295 accumulation of false-positive observations the specificity is poor and decreases with increasing  
296 numbers of surveys (Fig. 5).

## 297 **Discussion**


298 There can be few datasets where the same sites have been surveyed independently by such  
299 large numbers of people over such short periods. As the data were collected under exam  
300 conditions the independence of the surveys is largely assured, but also there is a clear incentive  
301 for participants to minimize their errors. The habitats, methods and participants are comparable

302 to surveys conducted routinely by professional and amateur surveyors. The surveys used in this  
303 study were largely conducted on grassland of various types, during the summer. Grassland is  
304 used in FISC assessments, because of its species richness and density, but also because it  
305 contains a wide variety of species that are known to be difficult to identify. These data are not  
306 particularly suitable for analysing habitat and seasonal dependent difference in surveying errors,  
307 but we believe that the conclusions drawn from these data are also applicable to botanical  
308 surveying in general. Moreover, many of the conclusions are likely to be applicable to surveys of  
309 other organisms, particular sessile organisms where identification is critical.

310 Botanical surveyors vary considerably, both in their ability to detect species and in their degree  
311 of caution (Fig. 1). In the surveys presented here the participants had a clear disincentive to  
312 avoid false-positive errors, whereas in many biological surveys there are in fact incentives to  
313 create false-positive errors. For example, amateur recording events such as bioblitzes and  
314 birding challenges incentivize the creation of long species lists. Having said that, Farmer et al.  
315 (2012) investigated the influence of incentives on errors in bird identification and found no effect.  
316 They did however, see widespread overconfidence among observers, which is consistent with  
317 large numbers of false-positive errors created by participants in these surveys.

318 False-positive errors could be found and corrected for if they occur predictably, *a priori* we  
319 expected that we would see a phylogenetic signal in false-positive errors. For example, it is well  
320 known that inexperienced botanists find the Poaceae and Cyperaceae difficult to identify.  
321 Furthermore, it is not unreasonable to expect recorders to mistake one species for its near  
322 relative. Yet, overall there is no clear phylogenetic signal to false-positive errors. So, phylogeny  
323 does not provide any solution to avoiding false-positive errors and it appears that errors are  
324 picked somewhat haphazardly from the pool of potential species. While observers may be less  
325 comfortable identifying difficult groups there is little evidence that more false-positive  
326 observations are created for these groups. Indeed, the difficulty of some groups could even

327 result in fewer false-positive observations, if their difficulty leads to greater caution by the  
328 observers. Nevertheless, the positive correlations of false-positive and cautious observations  
329 with the number of species in a genus indicate that errors are more likely for groups with similar  
330 identifying features.

331 Although  cautious records are collected in surveys they are rarely used in analyses. It is  
332 certainly preferable for surveyors to make cautious records, rather than making false-positive  
333 mistakes, however, we found no evidence that surveyors who made more cautious records also  
334 recorded fewer false-positive errors.

335 At all sites, regionally rare species are more likely than common species to be falsely observed.  
336 This trend is not restricted to plant identification, Farmer et al. (2012) found the same effect  
337 using simulated avian point count data. Knowledge of the regional abundance of species could  
338 be a useful tool for the automated validation of observations.

339 As expected, increasing the threshold for the acceptance of an observation increased the  
340 specificity of surveying at the expense of sensitivity. In this paper the acceptance threshold has  
341 been varied by requiring multiple observations for the acceptance of a species. The acceptance  
342 threshold for an observation will influence the specificity and sensitivity of that survey. However,  
343 there is no perfect threshold, and it should be chosen by the experimenter based upon the  
344 requirements of the experiment. For example, in comparison to using the results from a single  
345 survey, surveying a site three times and using an acceptance threshold of presence in two  
346 independent surveys will considerably improve the specificity without a loss in survey sensitivity.  
347 Nevertheless, this has been at the cost of three times as much surveying effort. In many cases  
348 setting such a high threshold for the acceptance of a species at a site is unacceptable, but,  
349 there may be occasions when such a threshold is appropriate. Furthermore, it is not necessary  
350 to set the same acceptance threshold for all species. Rare species, for example, are more likely  
351 to be recorded as false-positive observations, so they might need a higher threshold.

352 Furthermore, although we found no phylogenetic component to false-positive observations,  
353 there was considerable variability between species with respect to their likelihood of being a  
354 false-positive. If it were known which species are most susceptible to these errors then the  
355 threshold for accepting them could be greater.

356 There are other methods for increasing the threshold for acceptance that can be used by  
357 observers in the field. Systems for data entry that give feedback to the user on the likelihood of  
358 their observation would give users a chance to reflect on the data they have entered, thus  
359 avoiding common slips. Furthermore, IT systems that can support field identification with  
360 illustrations and keys, would avoid knowledge-based errors, particularly if such a system was  
361 geographically aware. Certainly, at least in Europe, considerable data on land cover, vegetation  
362 classification and species distributions exist, which, if it were available to field recorders, could  
363 reduce their chances of mistakes.

364 Were systems for automatic data validation available for field data collection then some  
365 predictable errors could be prevented. Errors such as non-existent species are completely  
366 preventable, but alerts for locally rare species would also reduce errors. Furthermore, errors due  
367 to an incorrect locality or date can be largely eliminated. Errors of common species are still  
368 harder to resolve, but simple slips can be validated by using feedback to the surveyor,  
369 particularly if there are already data available from the site for comparison.

370 Slips and lapses are common types of subconscious errors which, upon review, the person would  
371 know is wrong (Rasmussen, 1983). An example of a slip is writing down one species when a  
372 similarly named species was intended (e.g. *Carex divisa* instead of *Carex divulsa*) or the  
373 transposition of the numbers in coordinates. An example of a lapse is the omission of important  
374 information, such as writing down "*P. vulgaris*" in a situation where "*P.*" might refer to either  
375 *Pinguicula*, *Polygala*, *Primula*, *Prunella* or *Pulsatilla*.

376 Rules-based errors happen when people are working in familiar situations and applying routine  
377 rules that they have learned as part of their work, but these rules can either be unsuitable for all  
378 occasions, or can be misapplied. In the case of biological recording they may occur when  
379 people survey in areas they do not know well, and apply rules of thumb that work in their familiar  
380 area, but not in other areas where a different suite of species are present. Mistakes such as  
381 these would be reduced by facilitated access to keys in the field, particularly highlighting  
382 distinguishing characters.

383 Knowledge-based rules are where people are confronted with unfamiliar situations and try to  
384 apply their knowledge to come to a solution using more in-depth reasoning than simply applying  
385 rules. However, they may not appreciate the limits of their own knowledge and are unaware of  
386 the potential for error. Such errors come about due to overconfidence and can be compounded  
387 by confirmation bias. Wolf et al. (2013) show that people working together or at least with the  
388 knowledge of other people's decisions will create fewer false-negatives and fewer false-  
389 positives. Senders & Moray (1991) succinctly stated that "*The less often errors occur, the less*  
390 *likely we are to expect them, and the more we come to believe that they cannot happen*". Thus  
391 biological recording methodologies that fail to inform the surveyors of their errors are destined to  
392 add to the problem by contributing to their complacency.

393 Designing botanical surveys and analysing their results is particularly difficult due to the large  
394 number of species involved, which vary in their rarity and detectability. It is practically impossible  
395 to optimise methods for all species simultaneously and even optimising to the average ignores  
396 the fact that the species at either extreme of rarity are likely to be the ones we are most  
397 interested in. MacKenzie & Royle (2005) analysed the efficiency of survey design and  
398 concluded that rare species are more efficiently surveyed with extensive surveys, with less  
399 intensity at each survey site, whereas efficient surveying for common species was best under  
400 intense surveying of fewer sites. Unfortunately, extensive surveying for rare species, without

401 systems to reduce false-positive observations, will create more false-positive observations,  
402 reducing the benefits of increased efficiency. Intensive surveying of a few sites for common  
403 species is unlikely to be problematic, because common species are rarely false-positives.  
404 MacKenzie & Royle (2005) also found that there was no efficiency gain in double-sampling,  
405 compared to single surveys, though they did not consider the potential advantages of reducing  
406 false-positive observations.

407 In summary, false-positive observations are a common and pervasive problem in botanical  
408 surveying. The fact that they occur should not be ignored and when field survey methods are  
409 designed, they should be considered. The use of electronic data entry systems in the field  
410 should be promoted as their potential to validate, prompt and guide surveyors will reduce errors.

#### 411 **Acknowledgements**

412 The authors acknowledge the help of Gordon Leel and Mark Duffell who did the digitization.

#### 413 **References**

414 Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D.,  
415 Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N. & Frusher, S. (2014)  
416 Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*,  
417 173, 144-154.

418 Chapman, A.D. (2005) Principles of data quality. GBIF.

419 Chen, G., Kery, M., Zhang, J. & Ma, K. (2009) Factors affecting detection probability in plant  
420 distribution studies. *Journal of Ecology*, 97, 1383-1389.

421 Chen, G., Kéry, M., Plattner, M., Ma, K. & Gardner, B. (2013) Imperfect detection is the rule  
422 rather than the exception in plant distribution studies. *Journal of Ecology*, 101, 183-191.

- 423 Conn, P.B., McClintock, B.T., Cameron, M.F. Johnson, D.S., Moreland, E.E. & Boveng, P.L.  
424 (2013) Accommodating species identification errors in transect surveys. *Ecology*, 94, 2607-  
425 2618.
- 426 Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of  
427 presence-only data. *Global Ecology and Biogeography*, 23, 1472-1484.
- 428 Durka, W. & Michalski, S.G. (2012) Daphne: a dated phylogeny of a large European flora for  
429 phylogenetically informed ecological analyses. *Ecology*, 93, 2297.
- 430 Ellis, R. (2011) Jizz and the joy of pattern recognition: Virtuosity, discipline and the agency of  
431 insight in UK naturalists' arts of seeing. *Social Studies of Science*, 41, 769-790.
- 432 Elphick, C.S. (2008) How you count counts: the importance of methods research in applied  
433 ecology. *Journal of Applied Ecology*, 45, 1313-1320.
- 434 Farmer, R.G., Leonard, M.L. & Horn, A.G. (2012) Observer effects and avian-call-count survey  
435 quality: rare-species biases and overconfidence. *The Auk*, 129, 76-86.
- 436 Fritz, S.A. & Purvis, A. (2010) Selectivity in mammalian extinction risk and threat types: a new  
437 measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24, 1042-1051
- 438 Groom, Q.J. (2013) Estimation of vascular plant occupancy and its change using kriging. *New*  
439 *Journal of Botany*, 3, 33-46.
- 440 MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and  
441 allocating survey effort. *Journal of Applied Ecology*, 42, 1105-1114.
- 442 McKelvey, K.S., Aubry, K.B. & Schwartz, M.K. (2008) Using anecdotal occurrence data for rare  
443 or elusive species: the illusion of reality and a call for evidentiary standards. *BioScience*, 58,  
444 549-555.

- 445 Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011)  
446 Improving occupancy estimation when two types of observational error occur: non-detection and  
447 species misidentification. *Ecology*, 92, 1422-1428.
- 448 Molinari-Jobin, A., Kéry, M., Marboutin, E., Molinari, P., Koren, I., Fuxjäger, C., Breitenmoser-  
449 Würsten, C., Wölf, S., Fasel, M., Kos, I., Wölf, M. & Breitenmoser, U. (2012) Monitoring in the  
450 presence of species misidentification: the case of the Eurasian lynx in the Alps. *Animal*  
451 *Conservation*, 15, 266-273.
- 452 Orme, C.D.L. (2012) *The caper package: comparative analyses in phylogenetics and evolution*  
453 *in R*. URL <http://caper.r-forge.r-project.org> [accessed 28 February 2016]
- 454 Rasmussen, J. (1983) Skills, rules and knowledge; signals, signs and symbols, and other  
455 distinctions in human performance models. *IEEE Transactions on Systems, Man and*  
456 *Cybernetics*, 13, 257-266.
- 457 Rich, T.C.G. & Woodruff, E.R. (1992) Recording bias in botanical surveys. *Watsonia*, 19, 73-95.
- 458 Revell, L.J. (2012) phytools: An R package for phylogenetic comparative biology (and other  
459 things). *Methods Ecology and Evolution*, 3, 217-223.
- 460 Sabbagh, K. (2001) *A Rum Affair. A True Story of Botanical Fraud*. Da Capo Press, New York.
- 461 Senders, J. & Moray, N. (1991) *Human Error: Cause, Prediction and Reduction*. Lawrence  
462 Erlbaum, New Jersey.
- 463 Scott, W.A. & Hallam C.J. (2003) Assessing species misidentification rates through quality  
464 assurance of vegetation monitoring. *Plant Ecology*, 165, 101-115.
- 465 Simons, T. R., Audredge, M.W., Pollock, K.H. & Wettroth J.M. (2007) Experimental analysis of  
466 the auditory detection process on avian point counts. *Auk*, 124, 986-999.



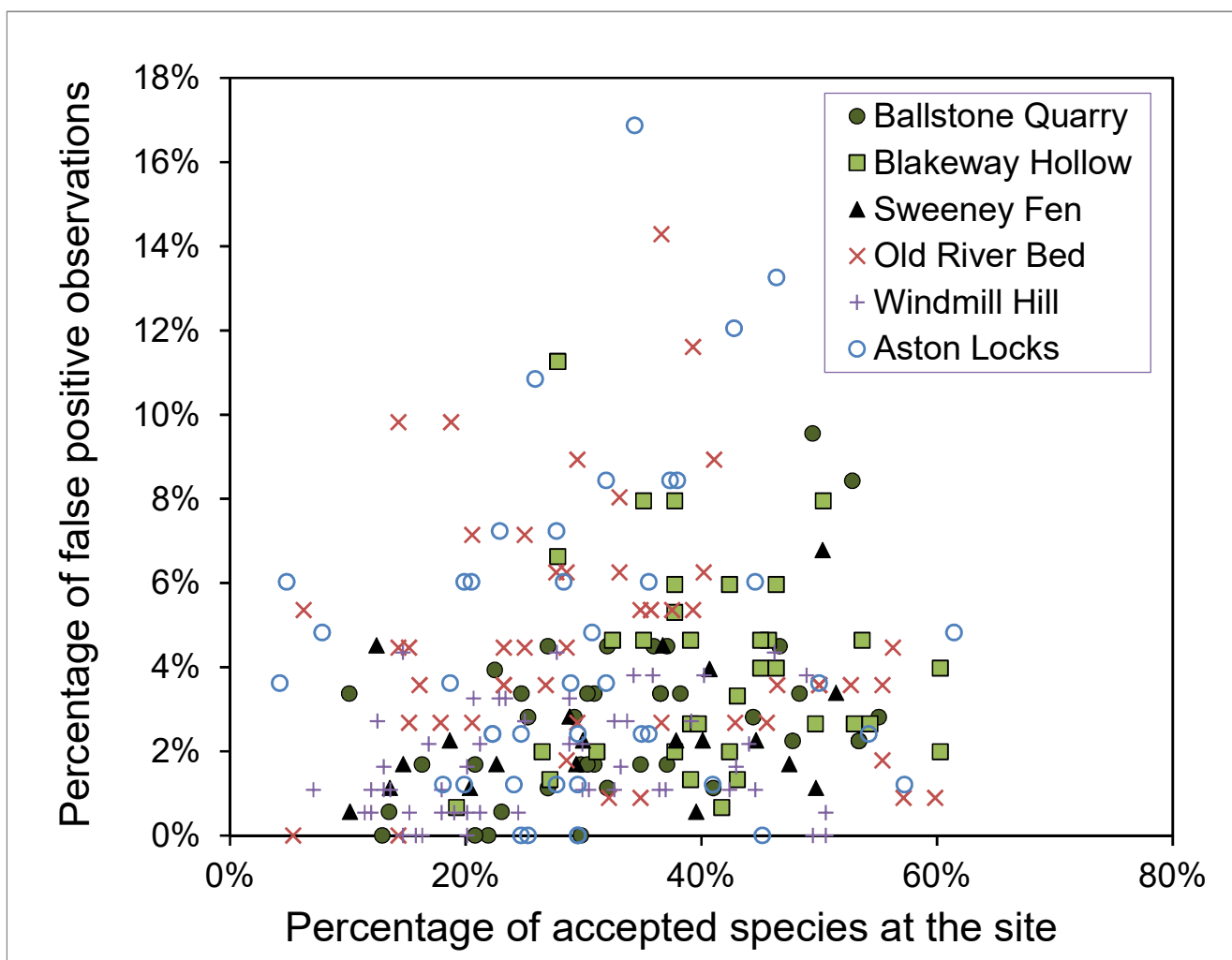
- 467 Stace, C. (2010) *New Flora of the British Isles*, 3rd ed. Cambridge University Press, Cambridge.
- 468 Willson, J.D., Winne, C.T. & Todd, B.D. (2011) Ecological and methodological factors affecting  
469 detectability and population estimation in elusive species. *The Journal of Wildlife Management*,  
470 75, 36-45.
- 471 Wolf, M., Kurvers, R.H.J.M., Ward, A.J.W., Krause, S. & Krause, J. (2013) Accurate decisions in  
472 an uncertain world: collective cognition increases true positives while decreasing false-positives.  
473 *Proceedings of the Royal Society of London B: Biological Sciences*, 280, 20122777.

**Figure 1** (on next page)

The percentage of false-positive versus correct observations for each observer.

The percentage of false-positive observations plotted against the percentage of correct observations from the list of accepted species at the site.

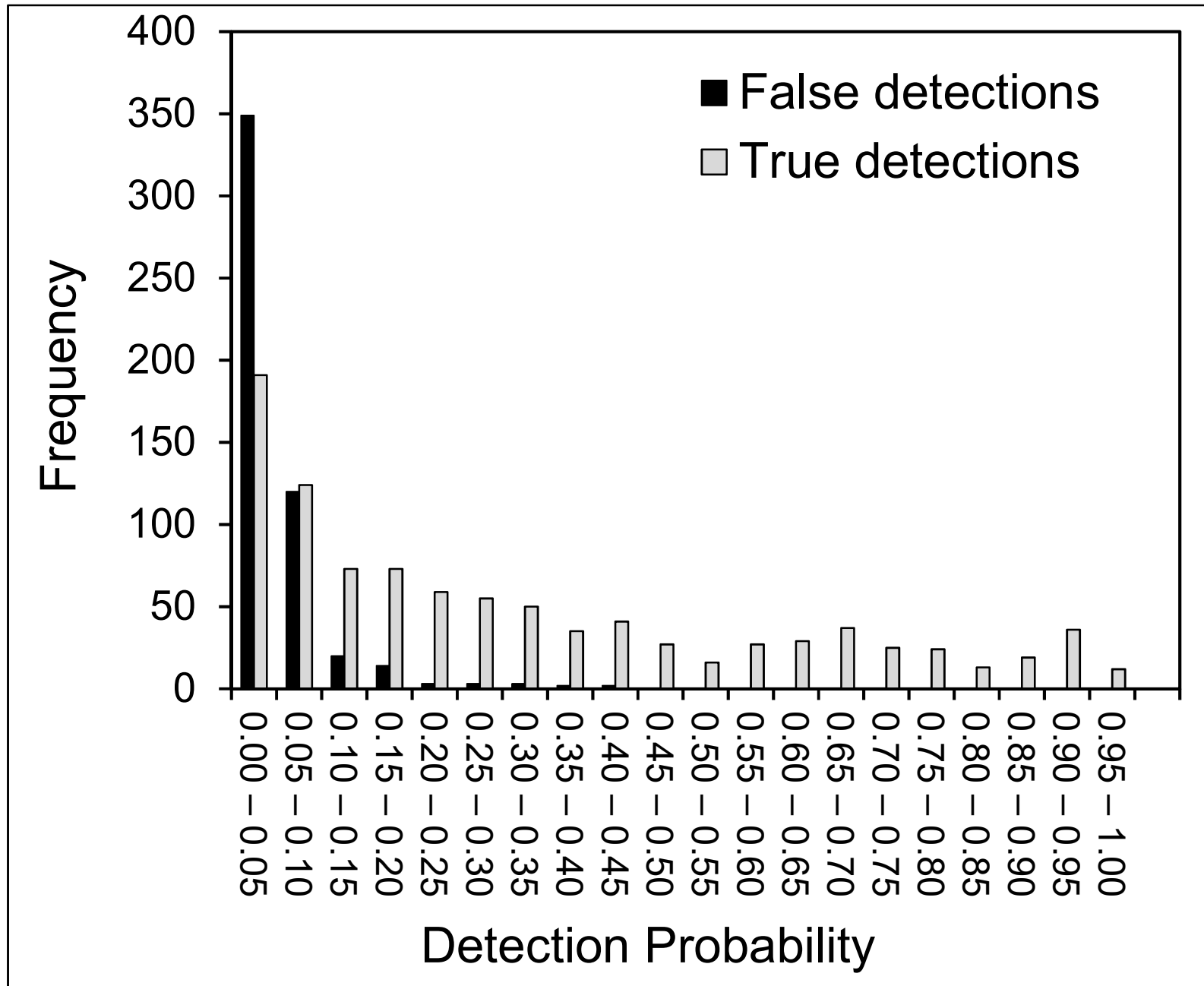




**Figure 2** (on next page)

The variation of detection probability between different plant species.

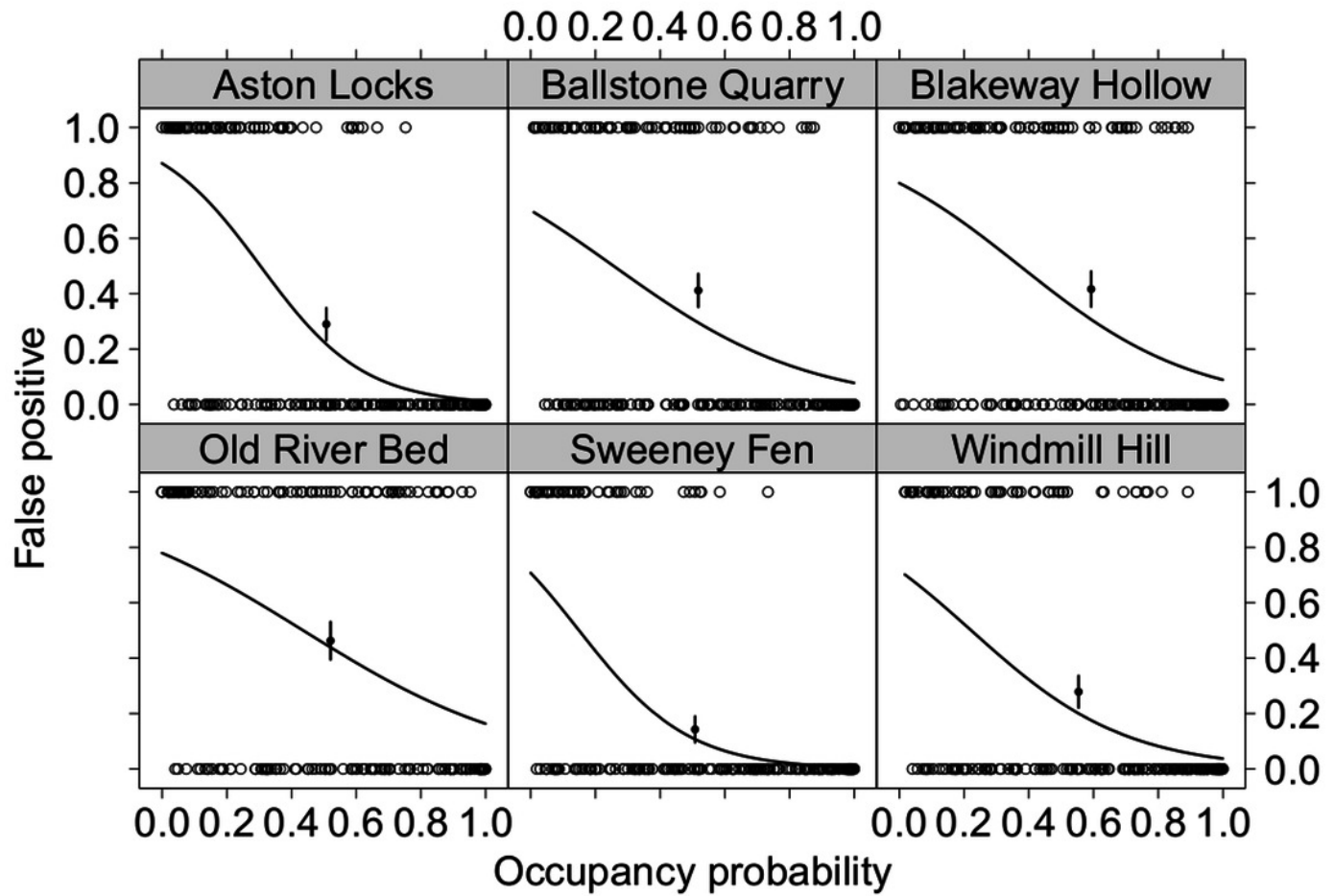
The distribution of detection probabilities for true observations and false-positive observations. Detection probability is calculated from the number of detections divided by the number of surveys.



## Figure 3



The relationship between the commonness of a species and the likelihood of it being observed when it is not there.

False-positive and true-positive observations for each species modelled as a function of the 4 km<sup>2</sup> occupancy probability of these species in southern England. False-positives were modelled as a binary response variable. To evaluate the goodness of fit of these models the proportion of false-positive errors was calculated from the central third of the data. It is shown with the point on the graphs together with its standard error. All model parameters are highly significant  $p < 0.01$ . The back transformed coefficients and their 95% confidence intervals where Aston Locks, intercept 0.871 [0.787 – 0.928], coefficient  $1.89 \times 10^{-3}$  [ $3.57 \times 10^{-4}$  –  $7.91 \times 10^{-3}$ ] d.f. 240; Ballstone Quarry, intercept 0.701 [0.579 – 0.803], coefficient  $3.48 \times 10^{-2}$  [ $1.17 \times 10^{-2}$  –  $9.16 \times 10^{-2}$ ] d.f. 246; Blakeway Hollow, intercept 0.800 [0.690 – 0.882], coefficient  $2.37 \times 10^{-2}$  [ $7.72 \times 10^{-3}$  –  $6.45 \times 10^{-2}$ ] d.f. 227; Old River Bed, intercept 0.780 [0.667 – 0.866], coefficient  $5.23 \times 10^{-2}$  [ $1.89 \times 10^{-2}$  –  $1.28 \times 10^{-1}$ ] d.f. 198; Sweeney Fen, intercept 0.708 [0.575 – 0.817], coefficient  $2.70 \times 10^{-3}$  [ $3.62 \times 10^{-4}$  –  $1.43 \times 10^{-2}$ ] d.f. 215; Windmill Hill, intercept 0.716 [0.586 – 0.822], coefficient  $1.52 \times 10^{-2}$  [ $4.02 \times 10^{-3}$  –  $4.88 \times 10^{-2}$ ] d.f. 229.

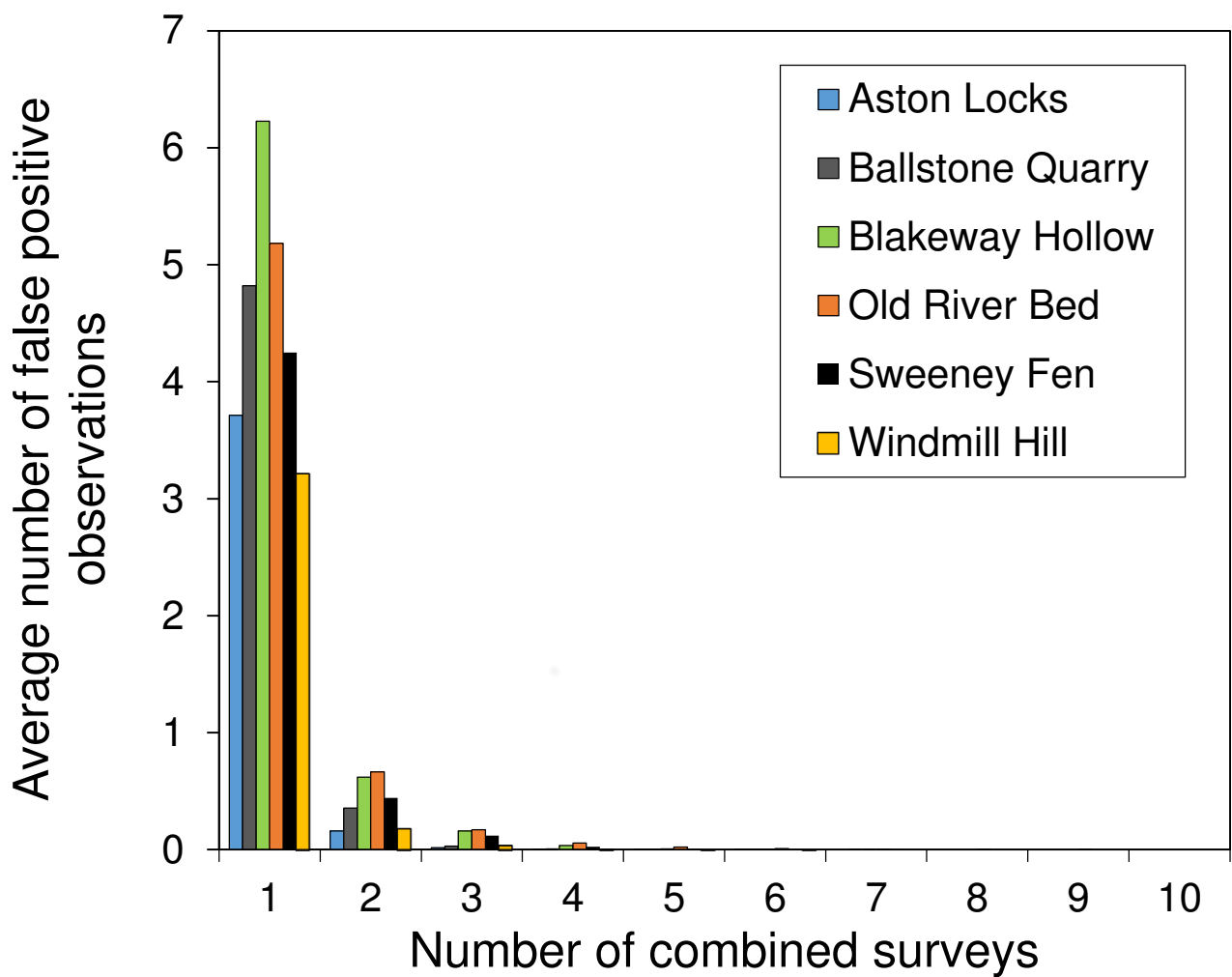


**Figure 4**(on next page)

Combining the results surveys dramatically reduced the number of false-positive observations.

The reduction in false-positive observations when the threshold for observation acceptance is increased. The graph shows the average number of false-positive observations with either one survey or two or more surveys  combined. Apart from when  is one survey, species are only accepted as present at a site if they are present in each survey.

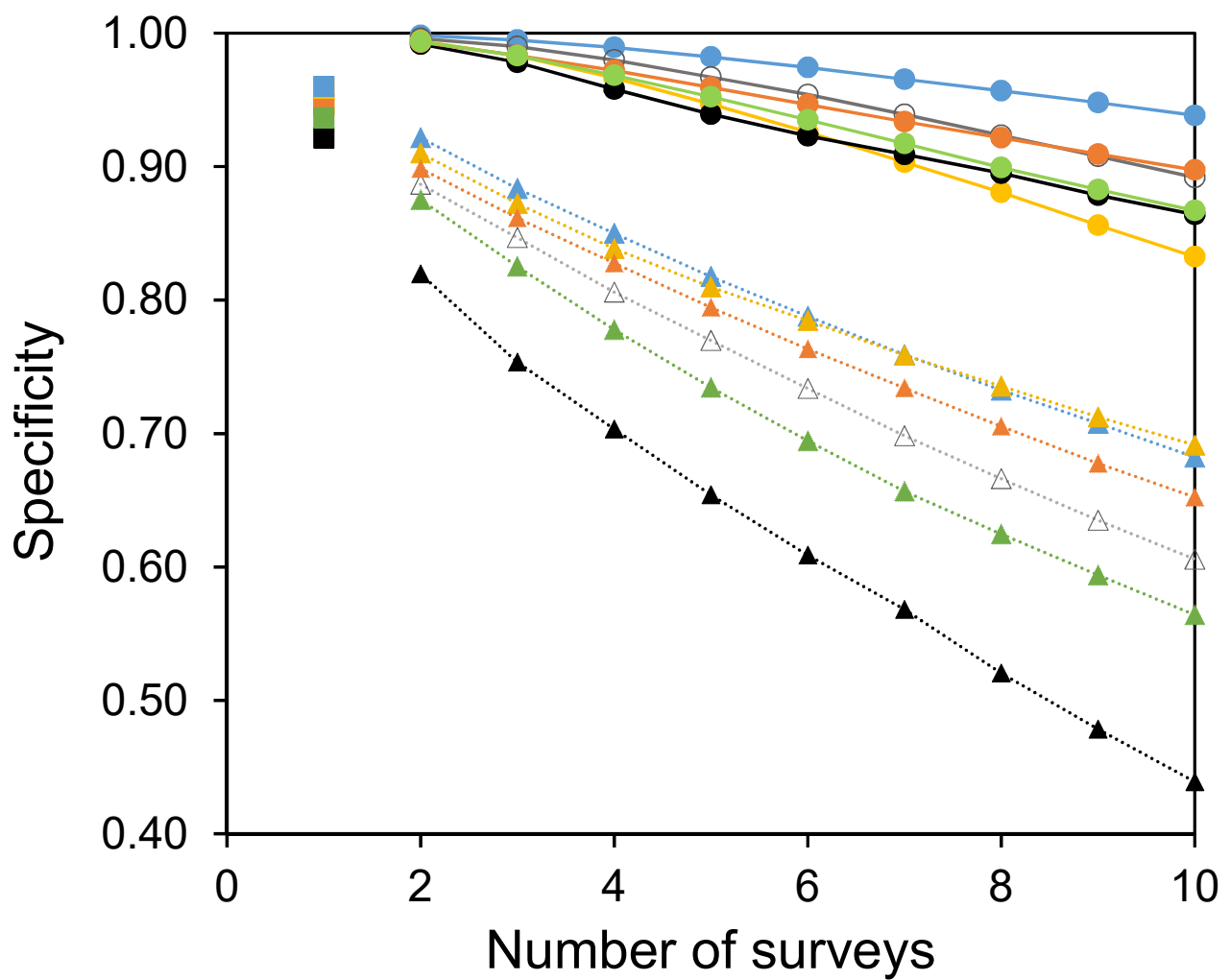




**Figure 5** (on next page)

The specificity of surveys at each site.

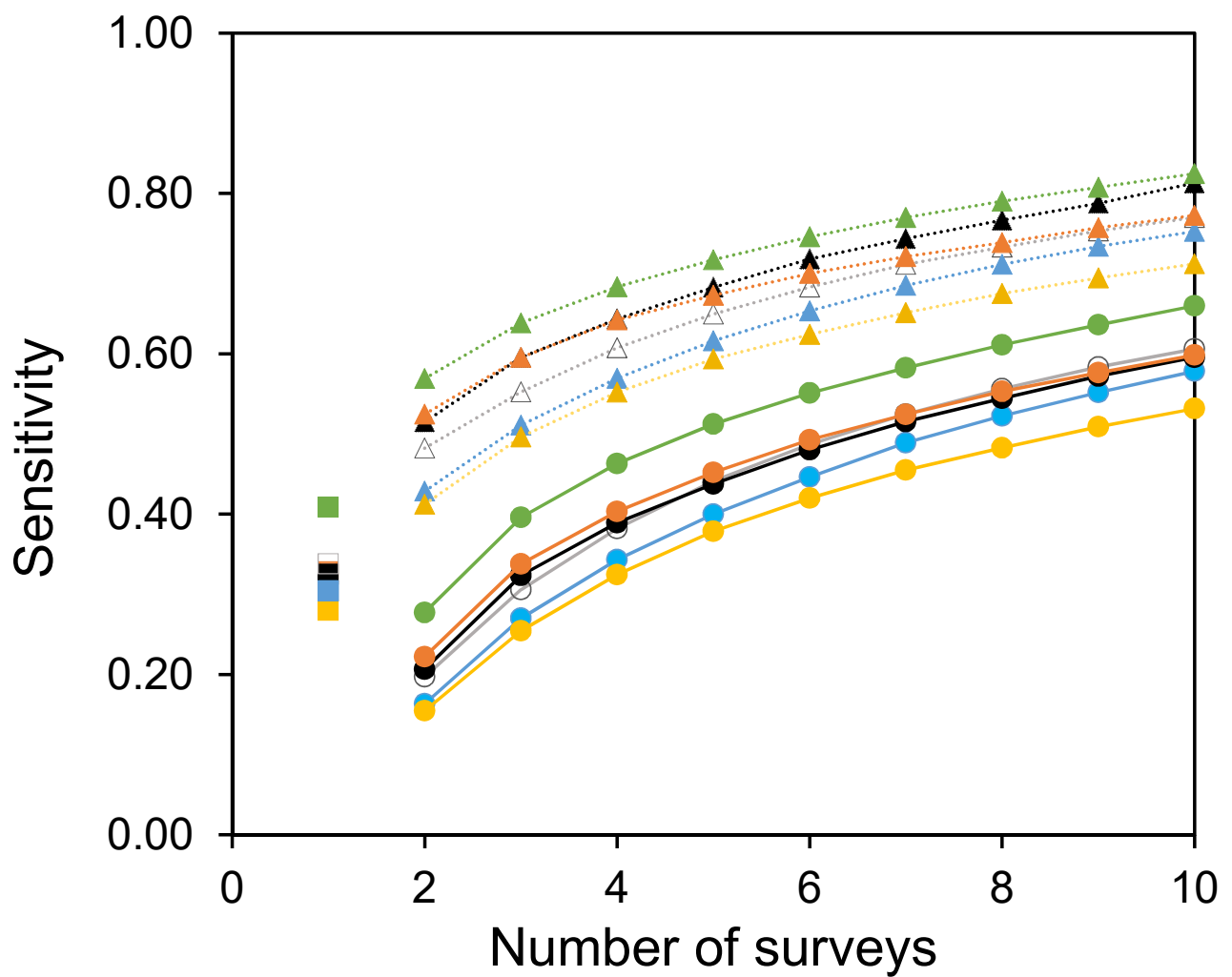
Circles with solid lines are where species are only accepted if they were found in two of the multiple surveys. Squares are where there was only one survey. Triangles with dotted lines are where all observed species are accepted. Colours represent the same sites as in figure 4. Specificity is calculated from the number of true negatives, divided by the sum of the true negatives and false-positives.



**Figure 6**(on next page)

The sensitivity of surveys at each site.

Circles with solid lines are where species are only accepted if they were found in two of the multiple surveys. Squares are where there was only one survey. Triangles with dotted lines are where all observed species are accepted. Colours represent the same sites as in figure 4. Sensitivity is calculated from the number of true positive observations, divided by the sum of the true positive observations and the false-negatives.



**Table 1** (on next page)

A summary of the survey sites.

The dates of the surveys and total numbers of surveys conducted, their location and habitat.

Site name	Survey dates	Number of surveys	Location (WGS84)	Habitat
Windmill Hill	27/06/2009 16/07/2009 01/07/2010	53	N 52° 36' 13" W 2° 33' 18"	grassland
Ballstone Quarry	04/07/2012 11/07/2012	35	N 52° 35' 35" W 2° 34' 35"	grassland
Blakeway Hollow	20/07/2011 31/07/2011	38	N 52° 35' 37" W 2° 34' 57"	grass verges and also dense species-rich hedges
Sweeney Fen	06/08/2007	20	N 52° 49' 4" W 3° 4' 38"	neutral grassland and calcareous fen
Old River Bed	24/06/2008 28/06/2008 14/09/2008	49	N 52° 43' 40" W 2° 44' 47"	sedge swamp
Aston Locks	02/07/2014 09/07/2014	42	N 52° 49' 51" W 2° 59' 18"	areas of open water, neutral grassland, tall herb and some sedge swamp

**Table 2** (on next page)

The phylogenetic signal strength of false-positive observations.

The D statistic for the species observed at each site, showing the phylogenetic signal strength of false-positive observations. The p value is the probability that the D value fits the model. Numbers in square brackets are the numbers of true-positive and false-positive observations in the sample.



	<b>Ballstone Quarry [173,86]</b>	<b>Blakeway Hollow [149,93]</b>	<b>Sweeney Fen [170,53]</b>	<b>Old River Bed [109,96]</b>	<b>Windmill Hill [177,71]</b>	<b>Aston Locks [162,91]</b>
D Statistic	1.107	1.064	0.916	1.088	0.992	1.112
<i>p</i> Random model	0.927	0.797	0.175	0.879	0.447	0.944
<i>p</i> Brownian model	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
<i>n</i>	259	242	223	205	248	253

1