

**A peer-reviewed version of this preprint was published in PeerJ on 17 May 2017.**

[View the peer-reviewed version](https://peerj.com/articles/3324) (peerj.com/articles/3324), which is the preferred citable publication unless you specifically need to cite this preprint.

Groom QJ, Whild SJ. 2017. Characterisation of false-positive observations in botanical surveys. PeerJ 5:e3324  
<https://doi.org/10.7717/peerj.3324>

# 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 observers 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 observers, some create a high proportion of false-  
24 positive errors, but these are scattered across all skill levels. Therefore, an observer's ability  
25 to correctly identify a large number of species is not a safeguard against the generation of  
26 false-positive errors. There was no phylogenetic pattern to falsely observed species,  
27 however, rare species are more likely to be false-positive as are species from species rich  
28 genera. In many studies the basic report of an organism is considered evidence of  
29 presence. However, raising the evidence threshold for the acceptance of an observation  
30 reduced false-positive observations dramatically, but at the expense of more false negative  
31 errors. False-positive errors are higher in field surveying of plants than many people may  
32 appreciate. Greater stringency is required before accepting species as present at a site,  
33 particularly for rare species. Combining multiple surveys from different observers resolves  
34 the problem, but requires a considerable increase in effort to achieve the same sensitivity as  
35 a single survey. Therefore, other methods should be used to raise the evidence threshold  
36 for the acceptance of a species. For example, digital data input systems that can verify,  
37 feedback and inform the user are likely to reduce false-positive errors significantly.

## 38 Introduction

39 Errors in science are inevitable. Sometimes they are the result of random chance, but also from  
40 human fallibility. Errors are particularly common in observations of biodiversity, where  
41 organisms can be either inconspicuous, hard to identify or hidden. Furthermore, the process of  
42 observation can be disrupted by external influences and observer biases (Simons et al., 2007;  
43 Willson et al., 2011). Animals are often intentionally secretive, but even sedentary organisms,  
44 such as plants, are difficult to observe owing to their similarity to each other. These sorts of  
45 errors lead to false-negative errors. False-negative errors are expected in plant surveys due to  
46 the variability in the detectability of different species in different seasons and habitats (Rich &  
47 Woodruff, 1992; Chen et al., 2009, 2013; Morrison, 2016). False-positive errors however, are  
48 those arising from observing something that was not there. These errors have been given many  
49 names, including detection errors, type 1 errors, errors of commission and misclassifications.  
50 Here we have chosen to use the terms false-positive and false-negative for the sake of  
51 readability. False-positive errors occur for several reasons; observers can either misidentify an  
52 organism; wrongly report the date or location or incorrectly transcribe otherwise correctly  
53 reported data.

54 While errors in vegetation surveying have been studied from many aspects there are few  
55 studies on false-positive errors in botanical recording and few survey schemes have specific  
56 quality assurance mechanisms, such as suggested by Scott & Hallam (2003). In contrast, much  
57 more attention has been paid to observation errors of animals where progress has been made  
58 in the methods for observation and analysis (Royle & Link, 2006; Elphick, 2008). Much expert  
59 plant identification of common taxa in the field is done using gestalt perception, rather than  
60 formal identification of characters (Ellis, 2011). However, human senses and reasoning, though  
61 remarkable, are prone to various sorts of error including apophenia (seeing apparently  
62 meaningful patterns within random data), generalizations and confirmation bias.

63 Users of botanical records expend considerable effort in “cleaning” data (Chapman, 2005).  
64 Cleaning entails using experience to verify and reformat the results of biological surveys,  
65 however, this is an inefficient process as errors are much better resolved close to their sources.  
66 Furthermore, data “cleaning” is also fallible, and elimination of errors early in the workflow is  
67 likely to be quicker and less costly than when time and distance is put between the observation  
68 event and the person analysing it. Statistical methods can also be used to account for observer  
69 errors and bias (Royle & Link, 2006; Miller et al., 2011; Bird et al., 2014; Dorazio, 2014), but  
70 while these approaches are useful, the first line of defence should be minimizing errors in field  
71 surveying.

72 In biological surveying, false-positive observations can be more costly than false-negatives. If  
73 false-negatives are suspected, additional surveying can help to resolve them. Also false-  
74 negatives are more likely to occur for rare species thus the more surveying conducted, the more  
75 one becomes confident that the species is either truly absent or at least extremely rare within  
76 your survey area. In the case of threatened species, false-negatives could lead to inappropriate  
77 actions in planning decisions or site management, but false-positives give the impression that a  
78 species is more common than it is and may lead to its conservation status not being recognised.  
79 The costs of errors needs to be assessed based on the goals of the survey, and should be  
80 considered in survey design and analysis methods (Fielding & Bell, 1997).

81 In contrast to false-negatives, false-positive errors are difficult to refute and can pollute datasets  
82 indefinitely. One only has to think of the time and effort wasted on extreme cases of false-  
83 positive errors, such as observations of plesiosaurs in Loch Ness and hominids in the Rocky  
84 Mountains, but there are many other examples (Sabbagh, 2001; McKelvey et al., 2008). Even in  
85 more benign cases Royle & Link (2006) showed, using a simulation, that ignoring false-positive  
86 observations in occupancy models could result in large biases in occupancy estimates.

87 In this study we use a quite unique set of plant surveying data where the same sites have been  
88 surveyed repeatedly by many independent observers. These data are derived from tests for  
89 Field Identification Skills Certificates (FISCs) that have been running for eight years under the  
90 aegis of the Botanical Society of Britain and Ireland. These certificates are intended to give the  
91 participants and potentially their future or current employers a guide to their skill at vascular  
92 plant identification in the field. These day-long tests include two laboratory-based tests and an  
93 afternoon field test and it is from this field test that the data in this paper are derived.

94 In classical signal detection theory, where the signal has to be separated from the noise, we can  
95 reduce the number of false-positive errors by increasing the detection threshold for accepting  
96 the signal. However, this is at the expense of an increase in false-negative errors (Wolf et al.,  
97 2013). We can examine this trade-off by changing the acceptance criteria for a species to be  
98 present. One simple method to increase the detection threshold is to combine the results of two  
99 or more observers, only accepting observations if multiple observers agree. So called double-  
100 observer methods are frequently used in animal surveys, particularly for avian point counts  
101 (Simons et al., 2007; Conn et al., 2013).

102 Double-observer methods can reduce the false-positive observations because false-positives  
103 are rare. If false-positives always account for a small proportion of the total number of  
104 observations and an observer picks their false-positive observations randomly from a fairly large  
105 pool of species names, then the number of false-positive observations of any one species  
106 should always be small and the chances of two observers picking the same false-positive  
107 species is extremely small.

108 However, there are two potential problems with this approach. Firstly, all the species that are  
109 actually present but only observed by one observer become false-negative observations.  
110 Secondly, the assumption that observers are unbiased in their creation of false-positive  
111 observations may not be true. For example, observers may pick false-positive observations from

112 species with similar character, such as those they are phylogenetically closely related to or they  
113 may pick them from common plants that they assume to be present.

114 In this paper we examine the characteristics of false-positive observations from the perspective  
115 of plant detectability, phylogenetic relatedness and their familiarity to observers. We examine  
116 whether changing the threshold for a true positive observation improves the accuracy of surveys  
117 and we discuss what other strategies could be used to reduce errors in botanical surveying. The  
118 intention is that the results can be used to design better plant survey methods that will lead to a  
119 reduction in the number of false-positive observations.

## 120 **Materials & Methods**

### 121 **Sites descriptions**

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

133 The site selection criteria were as follows.

- 134 • A relatively small more or less homogeneous site with fairly distinct boundaries.
- 135 • Small enough to survey thoroughly within two hours.



- 136       • Large enough for individual surveys to be carried out and for invigilation to be effective.  
137       • Possessing a reasonably complex vegetation with a good range of grasses, sedges  
138       and/or rushes, giving a total of around 100 vascular plant species to record.

139 Windmill Hill and Ballstone Quarry are grassland sites on Silurian Limestone. Blakeway Hollow  
140 is also on Silurian Limestone but is a sunken trackway with grass verges and also dense  
141 species-rich hedges. Sweeney Fen is a mixture of neutral grassland and calcareous fen over  
142 Carboniferous Limestone. The Old River Bed is an old meander filled with sedge swamp on  
143 Quaternary deposits. Aston Locks off-line reserve is adjacent to the Montgomery Canal and has  
144 areas of open water, neutral grassland, tall herbs and some sedge swamp.

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

148 Demographic data on observers were not collected, but anecdotally the age range is between  
149 25 and 60 with a median in the 30s. The gender balance is roughly equal. The main motivation  
150 for participants for taking a FISC has been career enhancement, with ecological consultants  
151 forming the bulk of participants. Occasionally, participants repeated the FISC, after gaining  
152 some experience in the field, however they were never tested on the same site twice.

153 The FISC field tests were evaluated by the number of correct species recorded in the field, and  
154 a score based on false-positive errors. Observers were clearly informed before the test that they  
155 would lose marks for false-positive observations giving them a clear disincentive to make  
156 mistakes. If they were not able to identify a plant to species they could report a genus for which  
157 they received half marks, but they could only do this once for a genus.

158

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

### 163 **Data**

164 The digitisation of the field data was carried out by a small team with instructions to pick out four  
165 categories of botanical records – correct species; unreasonable or known incorrect; 'mythical'  
166 species and 'cautious' records, with just a generic name. Correct species were those identified  
167 by the gold-standard observer, or species that had been recorded at the site by reliable  
168 recorders in other surveys of the site. Unreasonable species are those that have never been  
169 found at the site before. Mythical species were those that do not exist or do not occur in the UK.  
170 Data were simplified to species with taxonomy following Stace (2010), except for the  
171 phylogenetic signal tests where the taxonomy was aligned with the Daphne phylogeny (Durka &  
172 Michalski, 2012). The number of species for each genus was taken from Stace (2010), where  
173 species for each genus in the UK and Ireland are numbered.

174 Detection probabilities are calculated from the number of detections divided by the number of  
175 surveys.

176 The data used in this paper has been openly deposited in the Zenodo repository under DOI  
177 10.5281/zenodo.46662.

### 178 **Bootstrapping scripts**

179 The acceptance threshold for accepting a true-presence observation was varied by combining  
180 the results of two or more surveys. If the species was observed in each of the combined surveys  
181 then it was accepted as present. Therefore, the more surveys that are combined, the higher the  
182 threshold is to accept the presence of the species. Obviously, different surveys have varying

183 lists of observed species so to measure the average numbers of false-positive and false-  
184 negative observations when surveys were combined a bootstrapping approach was used.  
185 Depending on the number of surveys to be tested, a random selection of surveys was selected  
186 from the pool of all surveys conducted at each site. The number of false-positives and true  
187 positives were calculated from this selection and average values were calculated from 10,000  
188 randomly chosen selections. This script was written in Perl (ActivePerl, version 5.16.3.1603)  
189 and is available in a public Github repository together with a sample input file  
190 (<https://github.com/qgroom/grouped-surveys>).

### 191 **Statistics**

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

202 There are several tests for phylogenetic signal strength, particularly for continuous traits.  
203 However, as the majority of species had either one or zero false-positive observations we again  
204 chose to treat false-positives as a binary trait. To test for a phylogenetic component to the false  
205 detection of species the D statistic was calculated following Fritz & Purvis (2010). The D statistic  
206 is a measure of the phylogenetic signal strength of binary traits. The D statistic was calculated  
207 using the Caper package version 0.5.2 (Orme, 2012). The phylogeny used for the calculation

208 was the Daphne phylogeny of Northern European plants (Durka & Michalski, 2012). To  
209 calculate the D statistic for generic observations the Daphne tree was pruned using the Phytools  
210 package (Revell, 2012).

211 The specificity, in the context of botanical surveying, is the probability that a species is not seen  
212 given the plant is not present, it decreases with increasing numbers of false-positive  
213 observations and increases with the number of true-negatives (Fielding & Bell, 1997). The  
214 sensitivity is the probability that the plant is observed given that the plant is present, it  
215 decreases with increasing numbers of false-negative errors and increases with the number of  
216 true-positives. Mean specificity and sensitivity were calculated from the mean bootstrapped  
217 values for true positive, false-negative and false-positive observations. The value for true  
218 negative observations is somewhat arbitrary, because any number of species may not be  
219 present. So the number of true-negatives were calculated from the cumulative number of false-  
220 positive species recorded at the site, minus the average number of false-positive species in that  
221 survey or combination of surveys. This seems a reasonable value to use because the list of  
222 false-positive species are species of the British flora that were considered by the observers as  
223 being potentially present. Although this value is arbitrary, using other values only affects the  
224 absolute value of the specificity, not the relative values. Sensitivity is calculated from the  
225 number of true positive observations, divided by the sum of the true positive observations and  
226 the false-negatives. Specificity is calculated from the number of true negatives, divided by the  
227 sum of the true negatives and false-positives.

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

## 229 **Results**

230 The number of false-positive errors is not strongly related to the number of correct observations  
231 made (Fig. 1). Across all sites the number of false-positive errors is not significantly correlated

232 with the number of correct observations ( $R = 0.14$  [95% C.I.  $-0.11 - 0.40$ ,  $n=6$ ]). Observers,  
233 who can identify many species, are not necessarily likely to create few false-positive errors.  
234 Some observers are cautious while others are reckless.

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

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

242 The probability of a false positive observation occurring is always less than 0.5 and the majority  
243 of false detection probabilities are less than 0.05. Again, we should consider that the reference  
244 survey may have contained some false-negative errors, resulting in apparent false-positive  
245 errors for the observers. In such cases it is likely that several of the observers will have  
246 observed these species and this may explain the right-hand tail of higher false detection  
247 probabilities seen in Fig. 2.

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

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

253 Values of  $D$  usually vary between 0 and 1 and are inversely proportional to the degree of  
254 phylogenetic clustering. Values less than zero are expected if phylogenetically closely related  
255 species are more often identified as false-positive errors. Values greater than zero occur where

256 there is overdispersion, such that closely related species show opposite results. This would be  
257 the case where observers create errors that are closely related to the correct species. The D  
258 statistic is compared to the results from two evolutionary models, random and Brownian. Values  
259 close to 1 indicate that the random model is most appropriate and there is no phylogenetic  
260 relationship. The D statistic for false-positive observations at all six sites was close to one,  
261 indicating that there is no phylogenetic signal in the false detection of species (mean 1.04 [95%  
262 C.I. 0.98-1.11], n=6) (Table 2).

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

264 We use whether a species was chosen as a false-positive or a true positive as a binary  
265 independent variable and compared it with the average 4 km<sup>2</sup> occupancy for southern England.  
266 This allows us to compare how commonness and rarity relate to the likelihood of mis-  
267 observation in our surveys. For each site rarer species are more likely to be chosen in false  
268 positive observations (Fig. 3). From the y intercept of these models we can also estimate the  
269 probability of observing species that do not occur in southern England, which is 0.237 [95% C.I.  
270 0.184 – 0.291, n=6]; though this seems a large proportion, this value is cumulative for all  
271 surveys conducted at the site.

### 272 Inseparable taxa and fictional taxa

273 If observers were not able to identify a plant to species they could alternatively report a genus.  
274 Unlike errors these are conscious decisions to record a taxon at a less resolved taxon rank. The  
275 ten most recorded genera in this category were *Rosa*, *Salix*, *Euphrasia*, *Viola*, *Carex*,  
276 *Equisetum*, *Hypericum*, *Quercus*, *Epilobium* and *Myosotis*; listed in decreasing order of their  
277 number of observations. For all six sites there was a weak, but significant, positive correlation  
278 between the number of cautious observations and the number of species in a genus (R = 0.19  
279 [95% C.I. 0.27 – 0.10, n=6]). We did wonder whether observers who recorded more cautious

280 taxa would record fewer false-positives. However, there is no evidence for this in these data.

281 The number of cautious taxa recorded by an observer is not significantly correlated with their  
282 number of false-positive observations ( $R = 0.06$  [95% C.I.  $-0.19 - 0.07$ ,  $n=6$ ]).

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

286 In addition to cautious identifications there were non-existent taxa. These were a mixture of  
287 imaginary Latin and English names. They only made up a small number of observations, on  
288 average 3.7 observations per site. None were completely imaginary and most were  
289 combinations of words that exist in other names, but not together. For example, there were  
290 incorrect Latin combinations such as *Calystegia arvensis*, *Carex articulatus*, *Geranium palustre*,  
291 *Plantago ovalifolium* and *Polygonum vulgare*. Imagined vernacular names included black  
292 alkanet, burweed, separated rush and sheep's-beard. Names such as burweed do exist for  
293 species that do not occur in the British Isles, but it is assumed that the species intended by the  
294 observer was one that occurs in the British Isles with burred fruits, such as *Galium aparine*,  
295 *Geum urbanum* and *Torilis japonica*.

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

297 Increasing the threshold for acceptance of an observation dramatically reduced the number of  
298 false-positive observations (Fig. 4). Even when only two surveys are combined the average  
299 number of false-positive observations is 8% of the number in a single survey and false-positive  
300 observations are practically eliminated if five surveys are combined.

301 Two useful metrics for test performance are the specificity and sensitivity. Figures 5 and 6 show  
302 how the average specificity and sensitivity of plant observation change with different numbers of  
303 repeat surveys, using an acceptance threshold of presence in two surveys. With just two

304 surveys the specificity of surveying is close to its maximum value (Fig. 5). However, the  
305 increase in specificity has been at the expense of sensitivity, as combining two surveys creates  
306 many false-negatives (Fig. 6). Yet, with additional surveys the sensitivity rises quickly and  
307 exceeds the sensitivity of a single survey, whereas the specificity declines only slowly. With  
308 three surveys the sensitivity is approximately the same as a single survey, but the specificity  
309 has improved considerably.

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

## 315 Discussion

316 There can be few datasets where the same sites have been surveyed independently by such  
317 large numbers of observers over such short periods. As the data were collected under exam  
318 conditions the independence of the surveys is largely assured, but also there is a clear incentive  
319 for observers to minimize their errors. The habitats, methods and participants are comparable to  
320 surveys conducted routinely by professional and amateur surveyors. The surveys used in this  
321 study were largely conducted on grassland of various types, during the summer. Grassland is  
322 used in FISC assessments, because of its species richness and density, but also because it  
323 contains a wide variety of species that are known to be difficult to identify. These data are not  
324 particularly suitable for analysing habitat and seasonal dependent difference in surveying errors,  
325 but we believe that the conclusions drawn from these data are also applicable to botanical  
326 surveying in general. Moreover, many of the conclusions are likely to be applicable to surveys of  
327 other organisms, particular sessile organisms where identification is critical.



328 A potential criticism of the results is that there may have been errors in the surveys used as a  
329 baseline for the results. Nevertheless, we believe the number of these errors will be small and  
330 will not have biased the conclusions. We believe this because the results of more expert  
331 surveyors converge with the baseline data (Fig. 1). Also, there are no species seen by most of  
332 the observers, but not accepted as part of the baseline data. Furthermore, although errors in the  
333 baseline data could cause an inflation of the absolute numbers of false-positive observations,  
334 our conclusions are largely based on the relative number of false-positives, which would be  
335 unaffected by small errors in the baseline.

336 Botanical surveyors vary considerably, both in their ability to detect species and in their degree  
337 of caution (Fig. 1). In the surveys presented here the participants had a clear disincentive to  
338 avoid false-positive errors, whereas in many biological surveys there are in fact incentives to  
339 create false-positive errors. For example, amateur recording events such as bioblitzes and  
340 birding challenges incentivize the creation of long species lists. Having said that, Farmer et al.  
341 (2012) investigated the influence of incentives on errors in bird identification and found no effect.  
342 They did however, see widespread overconfidence among observers, which is consistent with  
343 large numbers of false-positive errors created by participants in these surveys.

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

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

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

361 At all sites, regionally rare species are more likely than common species to be falsely observed.  
362 This trend is not restricted to plant identification, the same effect also occurs in bird surveys  
363 (Manel et al., 1999; Farmer et al. 2012). Knowledge of the regional abundance of species could  
364 be a useful tool for the automated validation of observations.

365 As expected, increasing the threshold for the acceptance of an observation increased the  
366 specificity of surveying at the expense of sensitivity. In this paper the acceptance threshold has  
367 been varied by requiring multiple observations for the acceptance of a species. The acceptance  
368 threshold for an observation will influence the specificity and sensitivity of that survey. However,  
369 there is no perfect threshold, and it should be chosen by the experimenter based upon the  
370 requirements of the survey. For example, in comparison to using the results from a single  
371 survey, surveying a site three times and using an acceptance threshold of presence in two  
372 independent surveys will considerably improve the specificity without a loss in survey sensitivity.  
373 Nevertheless, this has been at the cost of three times as much surveying effort. In many cases  
374 setting such a high threshold for the acceptance of a species at a site is unacceptable, but,  
375 there may be occasions when such a threshold is appropriate. Furthermore, it is not necessary  
376 to set the same acceptance threshold for all species. Rare species, for example, are more likely  
377 to be recorded as false-positive observations, so they might need a higher threshold.

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

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

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

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

402 Rules-based errors happen when people are working in familiar situations and applying routine  
403 rules that they have learned as part of their work, but these rules can either be unsuitable for all  
404 occasions, or can be misapplied. In the case of biological recording they may occur when the  
405 observer is in an area they do not know well, and applies rules of thumb that work in their  
406 familiar area, but not in other areas where a different suite of species are present. Mistakes  
407 such as these would be reduced by facilitated access to identification keys in the field,  
408 particularly highlighting distinguishing characters.

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

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

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

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

#### 437 **Acknowledgements**

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

#### 439 **References**

440 Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D.,  
441 Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N. & Frusher, S. (2014)  
442 Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*,  
443 173, 144-154.

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

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

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

- 449 Conn, P.B., McClintock, B.T., Cameron, M.F. Johnson, D.S., Moreland, E.E. & Boveng, P.L.  
450 (2013) Accommodating species identification errors in transect surveys. *Ecology*, 94, 2607-  
451 2618.
- 452 Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of  
453 presence-only data. *Global Ecology and Biogeography*, 23, 1472-1484.
- 454 Durka, W. & Michalski, S.G. (2012) Daphne: a dated phylogeny of a large European flora for  
455 phylogenetically informed ecological analyses. *Ecology*, 93, 2297.
- 456 Ellis, R. (2011) Jizz and the joy of pattern recognition: Virtuosity, discipline and the agency of  
457 insight in UK naturalists' arts of seeing. *Social Studies of Science*, 41, 769-790.
- 458 Elphick, C.S. (2008) How you count counts: the importance of methods research in applied  
459 ecology. *Journal of Applied Ecology*, 45, 1313-1320.
- 460 Farmer, R.G., Leonard, M.L. & Horn, A.G. (2012) Observer effects and avian-call-count survey  
461 quality: rare-species biases and overconfidence. *The Auk*, 129, 76-86.
- 462 Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors  
463 in conservation presence/absence models. *Environmental conservation*, 24(01), 38-49.
- 464 Fritz, S.A. & Purvis, A. (2010) Selectivity in mammalian extinction risk and threat types: a new  
465 measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24, 1042-1051
- 466 Groom, Q.J. (2013) Estimation of vascular plant occupancy and its change using kriging. *New*  
467 *Journal of Botany*, 3, 33-46.
- 468 MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and  
469 allocating survey effort. *Journal of Applied Ecology*, 42, 1105-1114.

- 470 Manchester Metropolitan University (2007) Academic Regulations & Procedures Handbook:  
471 Academic Ethics. [http://www.mmu.ac.uk/academic/casqe/event/docs/academic\\_ethics.pdf](http://www.mmu.ac.uk/academic/casqe/event/docs/academic_ethics.pdf)  
472 (accessed 30 Apr. 2016)
- 473 Manel, S., Dias, J.M., Buckton, S.T. and Ormerod, S.J. (1999), Alternative methods for  
474 predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied*  
475 *Ecology*, 36: 734–747. doi:10.1046/j.1365-2664.1999.00440.x
- 476 McKelvey, K.S., Aubry, K.B. & Schwartz, M.K. (2008) Using anecdotal occurrence data for rare  
477 or elusive species: the illusion of reality and a call for evidentiary standards. *BioScience*, 58,  
478 549-555.
- 479 Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011)  
480 Improving occupancy estimation when two types of observational error occur: non-detection and  
481 species misidentification. *Ecology*, 92, 1422-1428.
- 482 Molinari-Jobin, A., Kéry, M., Marboutin, E., Molinari, P., Koren, I., Fuxjäger, C., Breitenmoser-  
483 Würsten, C., Wölfel, S., Fasel, M., Kos, I., Wölfel, M. & Breitenmoser, U. (2012) Monitoring in the  
484 presence of species misidentification: the case of the Eurasian lynx in the Alps. *Animal*  
485 *Conservation*, 15, 266-273.
- 486 Morrison, L. W. (2015). Observer error in vegetation surveys: a review. *Journal of Plant*  
487 *Ecology*, 9(4), 367–379.
- 488 Orme, C.D.L. (2012) *The caper package: comparative analyses in phylogenetics and evolution*  
489 *in R*. URL <http://caper.r-forge.r-project.org> [accessed 28 February 2016]
- 490 Rasmussen, J. (1983) Skills, rules and knowledge; signals, signs and symbols, and other  
491 distinctions in human performance models. *IEEE Transactions on Systems, Man and*  
492 *Cybernetics*, 13, 257-266.

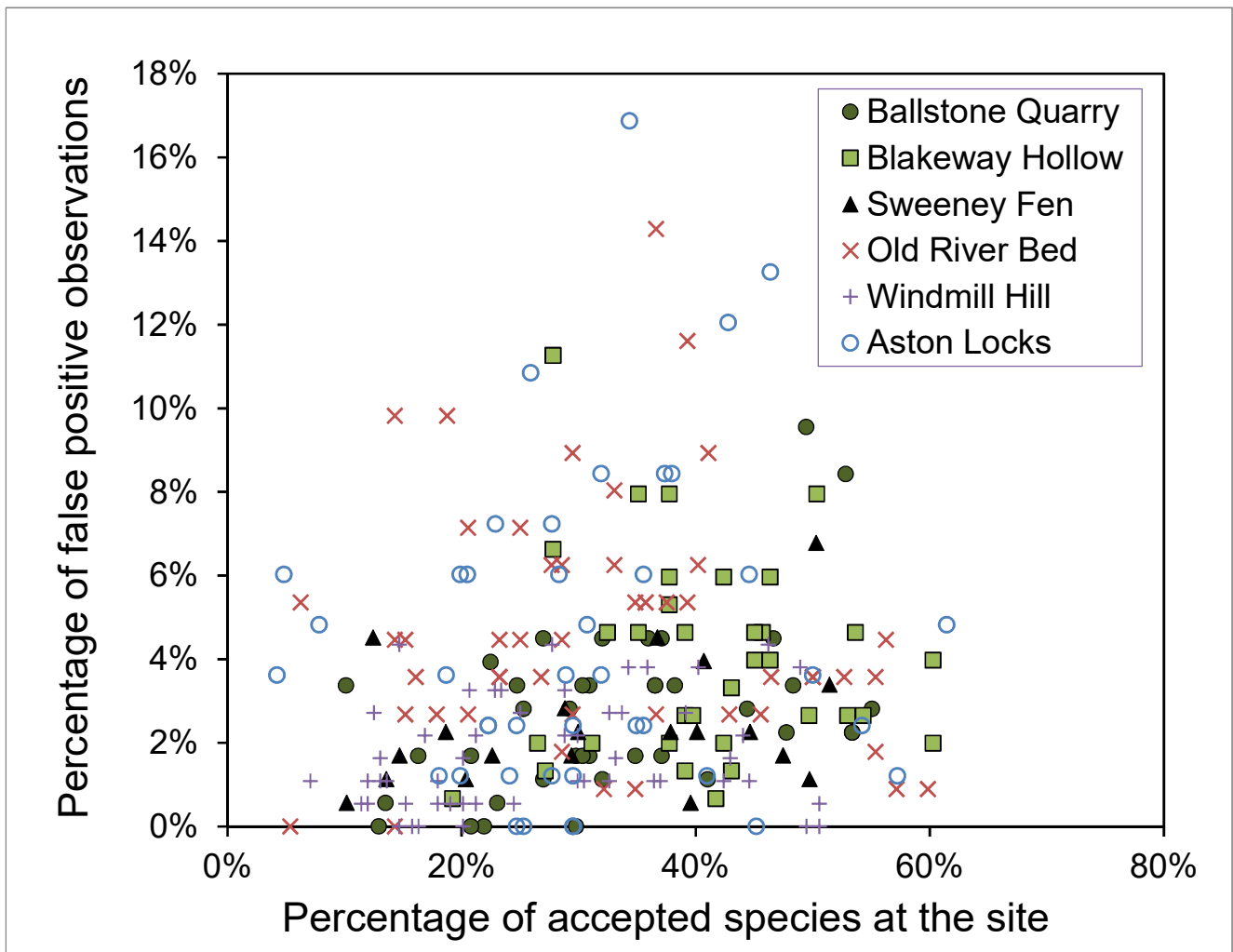
- 493 Rich, T.C.G. & Woodruff, E.R. (1992) Recording bias in botanical surveys. *Watsonia*, 19, 73-95.
- 494 Revell, L.J. (2012) phytools: An R package for phylogenetic comparative biology (and other  
495 things). *Methods Ecology and Evolution*, 3, 217-223.
- 496 Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive  
497 and false negative errors. *Ecology*, 87, 835-841.
- 498 Sabbagh, K. (2001) *A Rum Affair. A True Story of Botanical Fraud*. Da Capo Press, New York.
- 499 Senders, J. & Moray, N. (1991) *Human Error: Cause, Prediction and Reduction*. Lawrence  
500 Erlbaum, New Jersey.
- 501 Scott, W.A. & Hallam C.J. (2003) Assessing species misidentification rates through quality  
502 assurance of vegetation monitoring. *Plant Ecology*, 165, 101-115.
- 503 Simons, T. R., Audredge, M.W., Pollock, K.H. & Wettroth J.M. (2007) Experimental analysis of  
504 the auditory detection process on avian point counts. *Auk*, 124, 986-999.
- 505 Stace, C. (2010) *New Flora of the British Isles*, 3rd ed. Cambridge University Press, Cambridge.
- 506 Willson, J.D., Winne, C.T. & Todd, B.D. (2011) Ecological and methodological factors affecting  
507 detectability and population estimation in elusive species. *The Journal of Wildlife Management*,  
508 75, 36-45.
- 509 Wolf, M., Kurvers, R.H.J.M., Ward, A.J.W., Krause, S. & Krause, J. (2013) Accurate decisions in  
510 an uncertain world: collective cognition increases true positives while decreasing false-positives.  
511 *Proceedings of the Royal Society of London B: Biological Sciences*, 280, 20122777.



**Figure 1**(on next page)

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

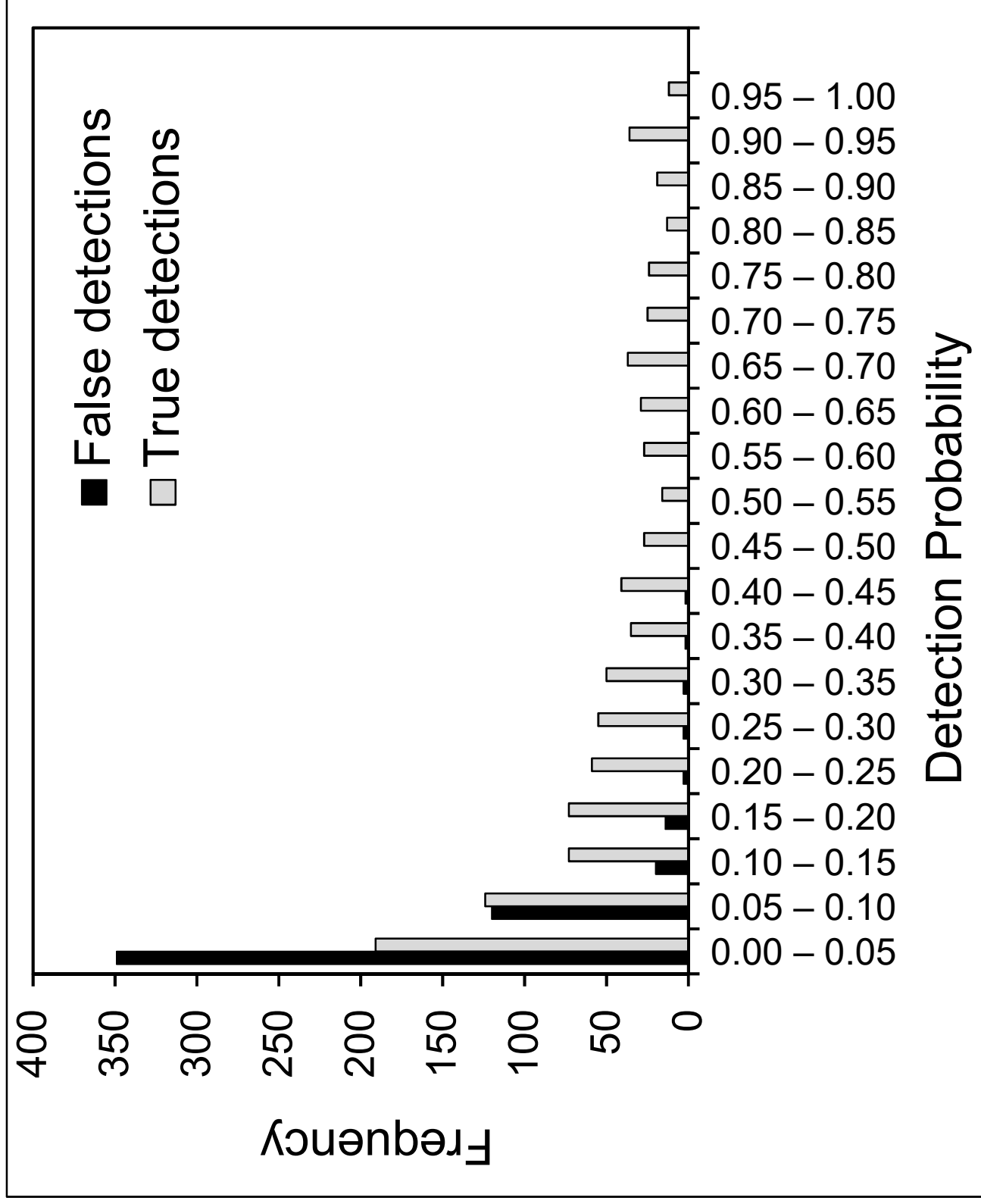
The percentage of false-positive observations plotted against the percentage of correct observations from the list of accepted species at the site created by 'gold-standard' observers and other reliable surveys.



**Figure 2** (on next page)

The variation of detection probability between different plant species.

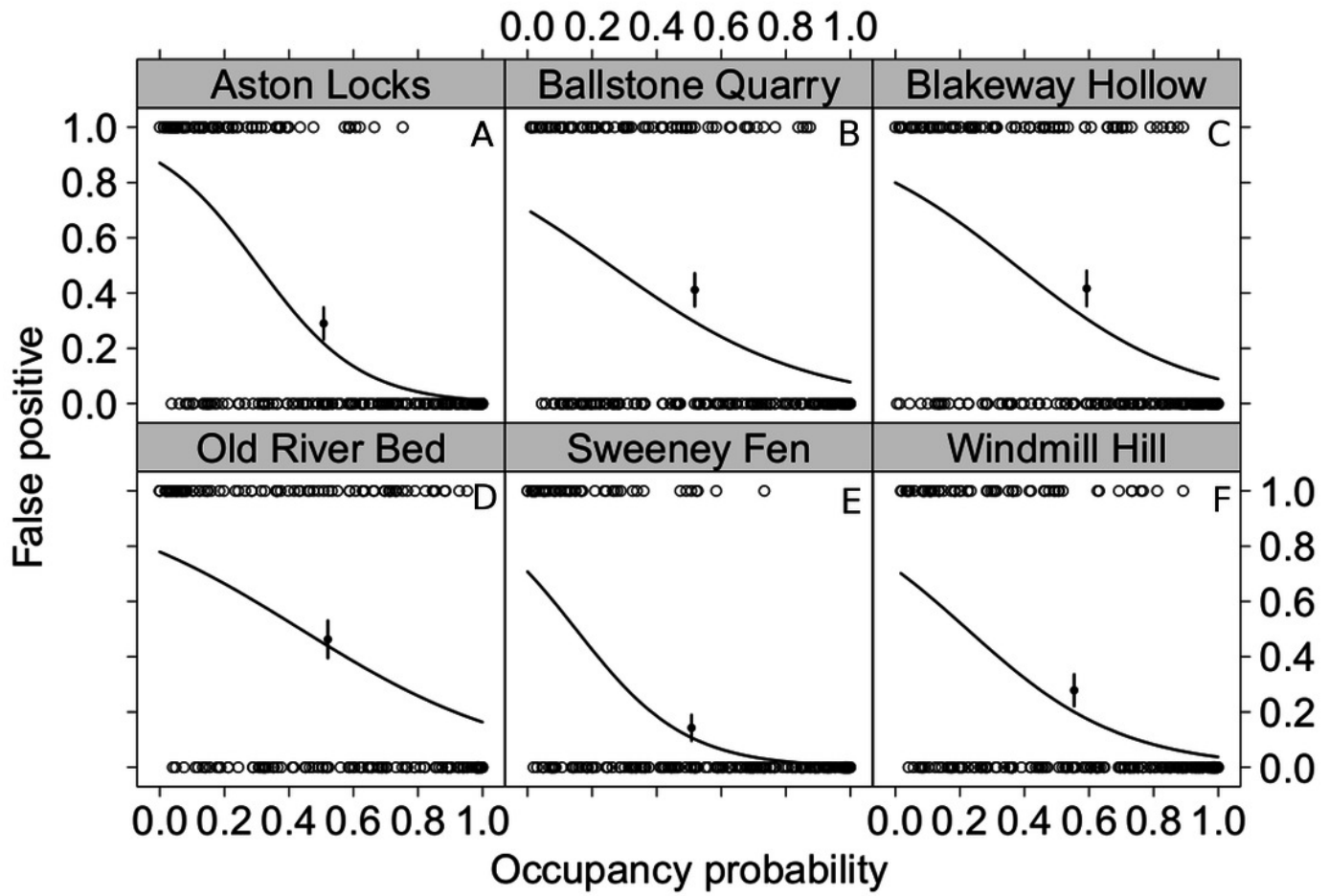
The distribution of detection probabilities for true observations and false-positive observations. Detection probability is calculated from the number of detections divided by the number of surveys.



## Figure 3

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

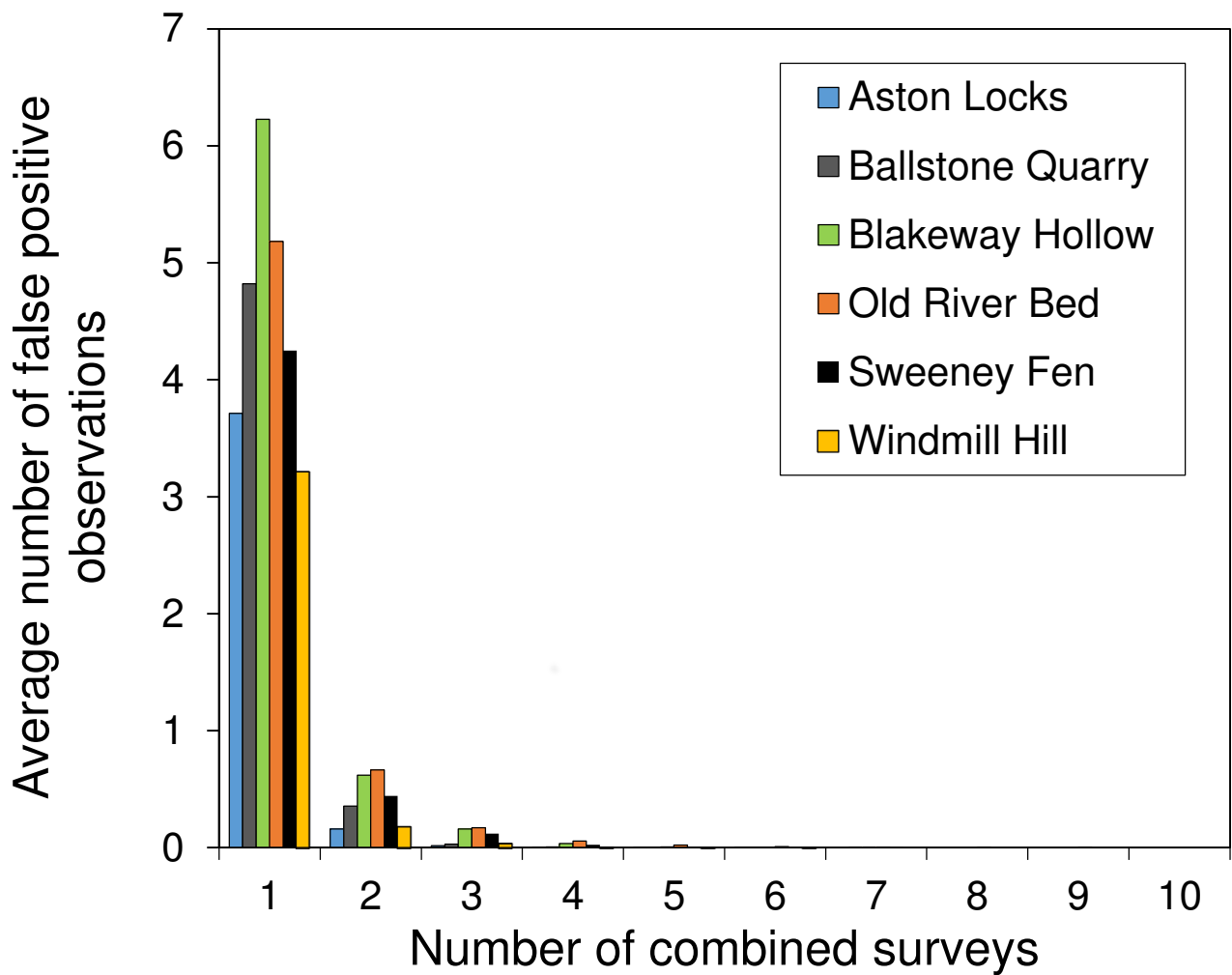
False-positive and true-positive observations for each species modelled as a function of the 4 km<sup>2</sup> occupancy probability of these species in southern England. False-positives were modelled as a binary response variable. To evaluate the goodness of fit of these models the proportion of false-positive 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 (A) 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; (B) 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; (C) 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; (D) 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; (E) 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; (F) 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 there is one survey, species are only accepted as present at a site if they are present in each survey.

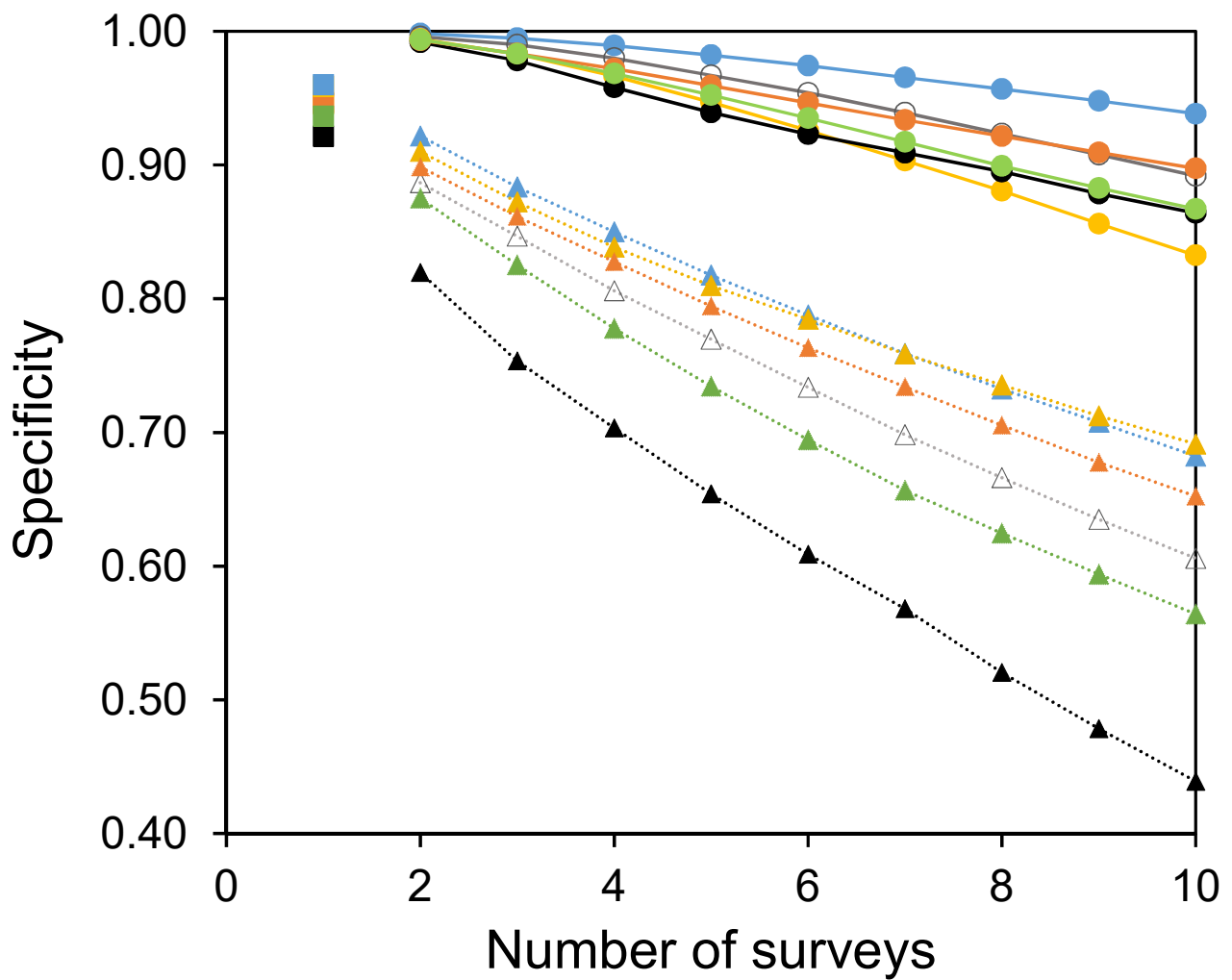




**Figure 5**(on next page)

The specificity of surveys at each site.

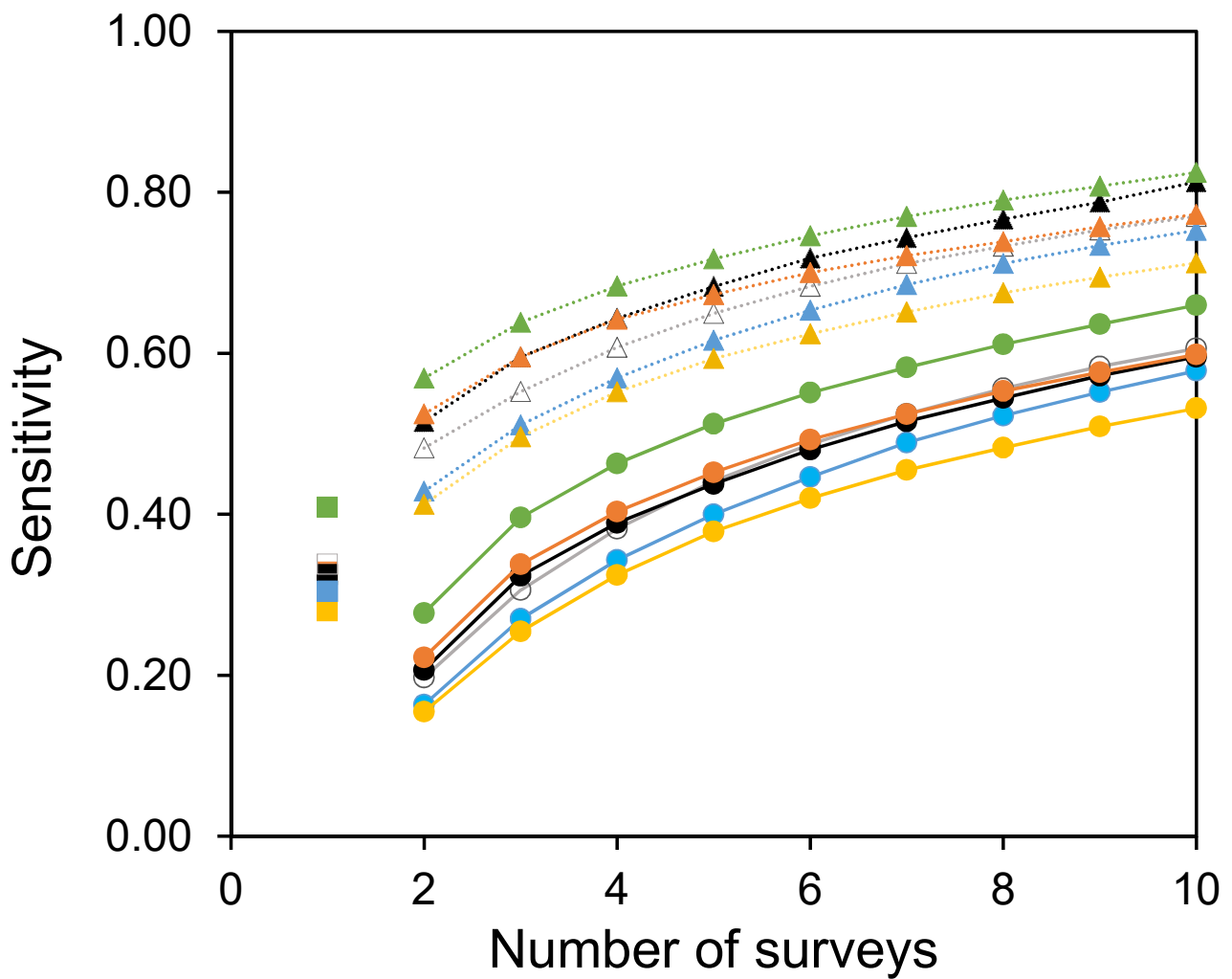
Circles with solid lines are where species are only accepted if they were found in two of the multiple surveys. Squares are where there was only one survey. Triangles with dotted lines are where all observed species are accepted. Colours represent the same sites as in figure 4. Specificity is calculated 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

1

**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
$p$ Random model	0.927	0.797	0.175	0.879	0.447	0.944
$p$ Brownian model	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$n$	259	242	223	205	248	253

1