Detection probabilities for sessile organisms

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>10.1002/ecs2.1546</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:29847623">http://nrs.harvard.edu/urn-3:HUL.InstRepos:29847623</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Detection probabilities for sessile organisms

Gabriele M. Berberich,¹, Carsten F. Dormann,² Dietrich Klimetzek,² Martin B. Berberich,³
Nathan J. Sanders,⁴ and Aaron M. Ellison⁵

¹University Duisburg-Essen, Faculty of Biology, Department of Geology, Universitätsstr. 5,
45141 Essen, Germany
²Albert-Ludwigs-University of Freiburg, Faculty of Environment and Natural Resources,
Department of Biometry and Environmental System Analysis, Tennenbacher Str. 4, 79085
Freiburg, Germany
³IT-Consulting Berberich, Am Plexer 7, 50374 Erftstadt, Germany
⁴Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark,
University of Copenhagen, Copenhagen 2100, Denmark
⁵Harvard University, Harvard Forest, 324 North Main Street, Petersham, Massachusetts, 01366
USA

†Corresponding author: Gabriele Berberich, email: gabriele.berberich@uni-due.de

Manuscript type: Article
Abstract
Estimation of population sizes and species ranges are central to population and conservation biology. It is widely appreciated that imperfect detection of mobile animals must be accounted for when estimating population size from presence-absence data. Sessile organisms also are imperfectly detected, but correction for detection probability in estimating their population sizes is rare. We illustrate challenges of detection probability and population estimation of sessile organisms using censuses of red wood ant (*Formica rufa*-group) nests as a case study. These ants, widespread in the northern hemisphere, can make large (up to 2-m tall), highly visible nests. Using data from a mapping campaign by eight observers with varying experience of sixteen 3600-m$^2$ plots in the Black Forest region of southwest Germany, we compared three different statistical approaches (a nest-level data-augmentation patch-occupancy model with event-specific covariates; a plot-level Bayesian and maximum likelihood model; non-parametric Chao-type estimators) for quantifying detection probability of sessile organisms. Detection probabilities by individual observers of red wood ant nests ranged from 0.31 – 0.64 for small nests, depending on observer experience and nest size (detection rates were approximately 0.17 higher for large nests), but not on habitat characteristics (forest type, local vegetation). Robust estimation of population density of sessile organisms – even highly apparent ones such as red wood ant nests – thus requires estimation of detection probability, just as it does when estimating population density of rare or cryptic species. Our models additionally provide approaches to calculate the number of observers needed for a required level of accuracy. Estimating detection probability is vital not only when censuses are conducted by experts, but also when citizen-scientists are engaged in mapping and monitoring of both common and rare species.
Key words:
Ants; citizen-science; detection probability; *Formica rufa*-group; Formicidae; Bayesian data-augmentation; non-parametric richness estimator; plot-level detection model; red wood ants; sessile organisms.
Introduction

Estimating population size is a central requirement of population and conservation biology. Similarly, estimating species ranges and predicting their changes – e.g. in response to climatic change and habitat disturbance – depends on accurately documenting presence and absence of individuals. In both cases, imperfect detection is a widely appreciated problem (e.g., Royle et al 2005, MacKenzie et al. 2006; Kellner and Swihart 2014; Dénes et al 2015): how can an observer be certain that individuals are detected when they are present? Consequently, estimates of detection probability now are used routinely in subsequent estimation of population sizes and ranges of common, rare, or cryptic mobile animals (e.g., Williams et al. 2011).

For sessile organisms such as plants, many marine invertebrates, and a wide range of colony-forming organisms including ants and termites, estimating their colony sizes or ranges would seem to be much easier than for animals that are constantly moving. However, detection probability of sessile organisms is surprisingly variable and strongly depends on the conspicuousness of the focal taxa; habitat characteristics; sampling design, time and duration; and the experience of the observer (e.g., Alexander et al. 1997, Miller and Ambrose 2000, Fitzpatrick et al. 2009). Sessile organisms also are simple targets for monitoring by citizen-scientists.

Ants are ubiquitous in most terrestrial landscapes (e.g., Dunn et al. 2009). Red wood ants (henceforth RWA) form very large, often polydomous colonies (Ellis and Robinson 2014); individual mound nests may reach 2 m in height and contain > 60,000 individual workers (Chen and Robinson 2013). RWA are of significant ecological importance (e.g., Klimetzek 1981, Way and Khoo 1992). Recently, RWA species have been introduced for biological control of undesirable insects (Seifert 2016), developed as biological indicators for otherwise undetected
tectonic activity (Berberich et al. 2016), and some are considered species of conservation concern (e.g., BfN 2012, IUCN 2015).

There are few long-term studies of RWA populations. Some investigators have suggested that populations of RWA are declining (e.g., Wellenstein 1990, Crist 2009), whereas others have reported that their populations are increasing (e.g., Stoschek and Roch 2006, Wilson 2011). Because none of these (or other) researchers have estimated or accounted for detection probability, a potential explanation for differences among studies is that estimates of occurrences or population sizes of RWA nests are inaccurate. Although this general problem has been recognized for mobile animals (e.g., MacKenzie et al. 2006), it is discussed only rarely in reviews of population sizes of endangered sessile species such as plants or ants (e.g., Philippi et al. 2001, Underwood and Fisher 2006, Godefroid et al. 2011). Therefore, we used the large, persistent, and highly apparent nests of red wood ants (*Formica rufa*-group) as a case study (Fig. 1).

Estimating the size of a population is a statistical problem addressed in hundreds of publications (e.g., Manning and Goldberg 2010, Grimm et al. 2014, Royle et al. 2015). Our case is different, although not atypical and several aspects render the application of established approaches either unnecessarily cumbersome or completely infeasible. First, as sessile organisms do not move, they do not have a capture or re-sighting history (as used, e.g., in Huggins-style recapture models, e.g. Akanda and Alpizar-Jara 2014): every time a plot is inspected, the nest will be found (with a certain detection probability) because the occupancy is constant (ψ = 1 for any object ever recorded). Second, detection probability is a function both of traits of the object (e.g., its size), and environmental conditions. Again, this has been addressed infrequently in recapture studies (but see Royle et al. 2004 for sparse data lacking object traits). The present
study employed several different statistical models, each of which is relatively simple and all of which estimate variability in detection rates by individual observers. An additional goal of the analysis was to quantify how many observers would be required to achieve a given level of accuracy for an estimator of population size. To achieve this goal, we also needed to estimate observer-specific detection probabilities.

In this study, we addressed five inter-related questions: (1) Do multiple observers detect or overlook the same RWA nest? (2) Is there a “best” way to quantify detection probability of sessile organism such as RWA nests? (3) Do colony size and density influence detection probability? (4) Does individual nest size influence detection probability? (5) How many observers are needed to converge on an estimate of the true number of nests? We asked these questions specifically with respect to individual RWA nests. In doing so, we improved estimates of RWA population sizes by including detection probability while simultaneously developing and using methods that will be applicable to a wide range of sessile organisms.

**Material and methods**

**Sampling design**

Field work was done during April 2015 in sixteen, randomly chosen 60 × 60-m plots near Friedenweiler (N47.54, E8.16, EPSG: 5677, 850 – 920 m a.s.l.) in the Black Forest region of southwest Germany. Eight observers (two experienced ones [co-authors GMB and MBB] and six inexperienced ones) independently mapped RWA nests for one hour in each of the 16 plots. The inexperienced observers were trained beforehand to recognize RWA nests in the field and to map them using a GPS receiver (Garmin 60CSx/62S/64S; 10-m precision) held directly above a RWA nest and register its location. Each observer also took a photograph of every mapped nest
(Fig. 1) to facilitate its subsequent identification and to avoid double-counting when nearby nests were within the precision of the GPS. Each GPS receiver was pre-loaded with 1:50,000 topographic maps onto which the boundaries of all 16 study plots had been transferred so that plot boundaries could be observed and maintained during each census.

All cameras and GPS receivers were synchronized to local time and projection (WGS84 projection; Datum: Potsdam). To avoid two observers mapping the same plot at the same time, each observer mapped the plots in a specifically defined sequence. The track of each observer in each plot was recorded continuously to quantify speed, total distance covered, and individual search strategy (Fig. 2). Finally, to minimize errors in delimiting plot boundaries in the field, a buffer region of 10 m around each plot was included during field recording to account for GPS imprecision. All GPS data were downloaded immediately after collection and transferred into a GIS database. Forest stand types were classified in the field, and nest heights and diameters were classified from nest photographs.

**Estimating and correcting for false positives**

False positives for each observer $i$ sampling in plot $s$ were tabulated manually from the number of reported nests. The number of observed real nests $N_{obs}$ was determined by cross-matching all mapped entities identified as RWA nests with their GPS coordinates, photographs, and recorded census tracks and expert knowledge. We linked GPS coordinate positions for each actual RWA nest recorded by each observer and averaged them to obtain a unique GPS position for each nest, which was then assigned a unique identifier. In all analyses, only real RWA nests were analyzed.
**Covariates of detection probability**

For exploratory analysis, we used a quasi-binomial generalized linear model to test whether nest sizes, classified by height-classes (1-10, 11-50, 51-100, and >100 cm) or diameter-classes (1-50, 51-100, 101-150 and >150 cm) of each ant nest (classified from nest photographs); the forest type (dominated by spruces \([Picea]\), pines \([Pinus]\) or beech \([Fagus]\)) in which it occurred (classified in the field); or its location (within the forest, along forest roads, or along forest edges, as classified in the field and from GIS layers) affected the number of nests detected by each observer. Because the number of small nests greatly exceeded those of larger nests, we pooled the two largest size classes when regressing detection probability on nest size.

**Statistical analyses**

Our data set is unusual relative to others in the detection-probability literature because (1) our objects do not move (in contrast to spatial recapture analyses, which estimate the probability of an animal having been observed in different plots, i.e., its occupancy); (2) we counted ant nests in several plots; (3) instead of plot-revisits (typical for recapture data), our “visits” were different observers, making it possible to determine observer-specific detection probabilities; and (4) each nest was characterized by its size, which may also have affected detection rates. Of course, there may be some nests that none of the eight observers discovered. For those we obviously also do not know the size or habitat characteristics.

We used three fundamentally different ways to estimate the total number \((\hat{N})\) of nests and the number of nests in each of our sampling plots, \(\hat{N}_s\).
Approach 1: Nest-level Bayesian data-augmentation

The most detailed analyses were done at the scale of individual nests (“nest-level” model). This nest-level model used a Bayesian data-augmentation approach to include the (potentially) overlooked nests in the analysis. For this analysis we used an approach similar to patch-occupancy models, which essentially included two elements. First, an indicator variable assigned each nest a value equal to 1 if it existed and to 0 otherwise. This indicator variable was drawn from a Bernoulli-distribution with a parameter representing the overall probability that a nest in the data actually existed. Second we used a logistic regression of the detection probability to account for observer-specific detection rates and effects of nest size and other covariates. The data (one row per nest) were augmented by 50 rows of missing data ($N_{\text{augmented}}$ unobserved nests, i.e., containing no information but contributing to the estimation of the overall probability that a nest existed; cf. Dorazoio et al. 2011). For the $N_{\text{augmented}}$ unobserved nests, the model estimated how likely it was that they were actually there, but were not observed. This could be achieved because the unobserved nests (and their sizes) were drawn from the same data model as were the observed data. The main tuning parameter of this nest-level model was the number of nests assumed to be missing; the model was insensitive to this parameter and yielded the same results when using 20, 50, or 200 augmented rows. Uninformative priors were chosen for all model parameters. The model was implemented in JAGS (Plummer 2003).

Approach 2: Plot-level detection models

We also estimated $\hat{N}_s$ using two different types of plot-level analyses: one Bayesian, and one using maximum likelihood. The disadvantage of these plot-level models is that they cannot accommodate nest-level information (e.g., size). On the other hand, the advantage of plot-level
models is that the maximum likelihood version can be used to readily simulate different numbers
of observers (requiring thousands of randomized analyses).

For each plot and for each observer, we modeled the number of nests observed as a
realization from a Binomial distribution, with parameters $\hat{N}_s$ and $\hat{P}_i$, representing the estimated
number of nests per plot $s$ and observer $i$’s detection rate, respectively: $P(\hat{N}_s, \hat{P}_i)$. Note that
this requires the estimation of $16$ (plots) + $8$ (observers) = $24$ different parameters. These
parameters could be estimated using Bayesian or maximum likelihood approaches, differing, in
our implementation, only in choosing (for the Bayesian version) priors for $\hat{N}_s$ that have a lower
bound at the observed number of nests at each plot. Then, for each plot × observer combination,
we estimated the expected number of observed nests as the product $\hat{N}_s \hat{P}_i$. As in the nest-level
model, we estimated a detection rate for each observer. Note that the Bayesian plot-level model
serves as a link between the data-augmentation model and the maximum-likelihood model,
illustrating that the main benefit of the data-augmentation approach is the incorporation of nest
sizes.

Finally, we used the maximum likelihood model to simulate estimates of nest counts that
we would get with fewer observers. To do so, we randomly drew $2$, $3$, ..., $7$ observers and re-ran
the estimation of nest numbers. Each simulation (number of observers) was repeated 1000 times.

**Approach 3: Non-parametric richness estimators**

Last, we used non-parametric sample-based estimators, developed for estimating the number of
species in samples of community data (Chao & Jost 2012; most recently reviewed by Chao et al.
2014). This approach does not account for observer-specific detection probability or plot-level
covariates. We estimated the total number of nests in each plot, $\hat{N}_s$, and the total number of nests
among the 16 plots, $\hat{N}$, using standard bias-corrected species richness estimators (Chao’s $S$, jackknife 1 [Jack1] $S$, and Jack2 $S$; see Chao and Jost 2012, Oksanen et al. 2015) implemented in the specpool function of the vegan library in R, version 3.2 (R Core Team 2015). These estimators are based on the observed number of nests that were detected by only one (“singletons”) or two (“doubletons”) observers.

**Determining the number of observers needed to accurately estimate the number of nests**

The analyses described above assumed that detection probability were independent of each observer. However, our data showed that many nests were recorded by all observers, whereas others were found only by some (Fig. 3). In other words, we could not assume independence of observations: adding more observers to the team led to records largely similar to what had already been reported. We computed the amount of effort required to accurately estimate numbers of nests assuming a constant detection probability among observers and serial correlation among observers (details of these calculations are given in Appendix S1).

Essentially, we estimated how more observers would affect our estimation, by assuming that new observers would have detection rates similar to those of our eight real observers, $P_i$. In addition to the detection rate of each observer, we had to compute the probability of a second observer finding a new nest, $P_c$, which we computed from the observed data for each observer pair. The probability that $k$ observers would overlook a nest was computed as $(1 - P_c)^k (1 - P_i)$. We simulated data for 9 and 10 observers, bootstrapping values for $P_c$ and $P_i$ based on our eight observers.
Availability of data and code

The commented R-code for all our analyses and figures are provided as online supplementary material (Appendix S2). All data and raw R and JAGS codes are available from the Harvard Forest Data Archive (http://harvardforest.fas.harvard.edu/data-archive), dataset HF-XXX.

Results

Sampling effort

Although the sampling protocol specified that each observer spend 60 minutes in a plot, GPS records revealed that actual time spent by the single observer in each plot ranged from 30 – 120 minutes. On the other hand, the eight observers were highly consistent in their searching behavior and all appeared to cover the majority of each plot in their searches while avoiding wetlands and very dense vegetation (Fig. 2). However, there was a surprising lack of consistency in the nests detected and overlooked by the different observers (Fig. 3).

Estimates of detection probability and the number of nests

Estimated detection probability ($\hat{P}_i$) computed from the nest-level model ranged from 0.37 to 0.64 (mean = 0.50). The plot-level models yielded estimates ranging from 0.31 – 0.52 (mean = 0.42; Bayesian plot-level detection model) or from 0.35 – 0.58 (mean = 0.47; plot-level maximum likelihood model; Table 1). Results of the Bayesian plot-level detection model suggested that we overlooked approximately 26% of nests (of an estimated total of 190 nests). The difference between the nest-level and plot-level estimates can be attributed to (1) fewer data points (the plot-level model aggregates all nests within a plot: $16 \cdot 8 = 128$ vs. 147 for the data augmentation); and (2) the joint estimation of detection rates and true number of nests,
P\left(N_{i,s}^{obs} \mid \hat{P}_i, \hat{N}_s\right)$, rather than conditionally $P\left(N_{i,s}^{obs} \mid \hat{N}_{i,s}\right)$, was estimated for each nest as in the data-augmentation model.

Estimated number of nests per plot ($\hat{N}_s$) ranged from 0 – 24 (patch-occupancy model), 0 – 27 (maximum likelihood) or 0 – 29 (Bayesian) (Table 2). Estimated total number of nests ($\hat{N}$) across all 16 plots = 147.7 (95%-confidence interval = [147, 149]), i.e. 1 to 3 nests overlooked; patch-occupancy model), 168.2 (maximum likelihood), or 190.1 (26% of nests overlooked; Bayesian). Estimated detection probabilities for the observers were slightly higher in the patch-occupancy model, but the estimated number of nests varied by a smaller percentage among models. In other words, while nest size affected detection probability, it did not greatly bias estimates of the total number of nests.

All of these estimates of total number of nests exceeded the bias-corrected ones that did not explicitly incorporate detection probability (Fig. 3; Table 3).

**Covariates of detection probability**

Large nests had a higher chance of being detected (estimate for $\beta_{size} = 0.819$). Height was a better predictor than diameter, making it necessary to incorporate nest height in an ideal analysis of these data. But nest size did not bias greatly estimates of the total number of nests. Detection probability increased significantly with both nest height (both linear [estimate = 6.7] and quadratic [estimate = -3.8] terms were significantly different from 0 [$P = 0.002$ and $P < 0.001$, respectively]) and diameter (only linear term [estimate = 1.3] was significantly different from 0 [$P < 0.001$]) (Fig. 4). Moreover, we found no relationship between the number of nests per plot and detection probability (Fig. 5). There also were no significant effects of forest type, position,
or interactions between these plot characteristics and nest-height size-class on nest detection
(Fig. 6 and Table 4).

**Effects of having more observers**

We observed that some of the 147 observed nests were detected by all observers (black squares in Fig. 3), whereas others were detected only by a single observer (white squares to the left of $N_{obs} = 147$ in Fig. 3). The average correlation among pairs of observers in detecting a nest was relatively high (0.65, $SD = 0.071$). Nonetheless, each new observer added some additional information. Assuming that still more observers would be similar to those we worked with, we found that there was an inverse relationship between the number of observers and $\hat{N}$: fewer observers led to higher estimates of overlooked and hence of the true number of nests (Fig. 7) because there are many nests but detection probability was relatively low. However, as the number of observers increased, fewer nests were overlooked (< 1% with eight observers; see Fig. 7, inset), and consistency among observers refined (and shrank) the estimated number of nests (Fig. 7).

**Discussion**

Our work with red wood ants addressed five general questions: (1) Do multiple observers detect or overlook the same nest? (2) Is there a “best” way to quantify detection probability? (3) Do colony size and density influence detection probability? (4) Does individual nest size influence detection probability? (5) How many observers are needed to converge on an estimate of the true number of nests? For RWA, the short answers are:

- Multiple observers detect and overlook different individual nests;
- Bayesian methods provide more precise estimates of detection probability;
• Population size and density had little effect on detection probability;
• Larger nests were more likely to be detected; and
• More observers are better, but the “return on investment” is a diminishing function.

Over the past several decades, a number of statistical models have been developed to correct for imperfect detection in population studies with respect to occupancy/species distribution modeling (reviewed in MacKenzie et al. 2006), mark-recapture (e.g. Lettink and Armstrong 2003, Chen and Robinson 2013) or distance sampling (Baccaro and Ferraz 2013). Many of these methods account for bias of observer, time of day, or season (Dénes 2015). Survey-, plot-, and species-level factors differentially affecting detection of species or individuals are incorporated only partially in these models, resulting in a disproportionately high number of non-detections (Iknayan et al. 2014, Dénes 2015). These issues are of particular concern for mobile organisms, but also can play a significant role for sessile ones (Chen et al. 2013). Additional difficulties also may arise when the objects under study vary in size or shape over time and are generally not easily noticed by unpracticed observers (e.g., Fitzpatrick et al. 2009).

The data-augmentation approach we used is fully in line with already published approaches (Royle et al. 2007, Kéry & Royle 2010, Dorazio et al. 2011). It models detection probability in exactly the same way, but the novelty is that it adds a characteristic for each individual nest, and it estimates the number of unobserved nests. There was only a small proportion of nests that were observed by only one observer, unlike e.g., the American redstart data in Royle (2004). One advantage of our patch-occupancy model with event-specific covariate (nest size) approach was that it allowed us to model each nest separately and thereby include a
The estimates of detection probability and nest abundance were similar between the Bayesian and maximum-likelihood models.

Our study of RWA nests highlights some underlying aspects of detection probability for sessile organisms. Red wood ants are ecologically important and have been listed as threatened or endangered because repeated censuses often suggest declines in abundance (e.g., Dekonick et al. 2010). However, detection probability of RWA nests has been estimated only once previously using a “mark–release–recapture method” while disturbing the ant colony (Chen and Robinson 2013). Our results, applying a non-invasive method without disturbing the ant colony, revealed that even in a well-designed survey of a well-known population, RWA nests were detected imperfectly even by experienced observers. Imperfect detection can bias seriously conventional estimators of species distributions and population sizes (Chen et al. 2013). Given a detection probability of RWA nests by experts of $\approx 0.63$, prior assertions of RWA decline (Dekonnick et al. 2010, IUCN 2015) should be revisited. Corrections for detection probability not only should be included in future inventories of RWA populations and other sessile organisms, but also should be accounted for in decisions to list these species as threatened or endangered.

Numerous covariates affect detection success (Dénes et al. 2015). Our results suggest that observer experience strongly influenced detection success of RWA, which also had been noted for other essentially sessile insects (Fitzpatrick et al. 2009). Whereas beginners and experienced observers both were highly consistent in their searching per plot, beginners identified fewer RWA nests. Experienced observers consistently detected twice as many short RWA nests (1-10 cm in height), observed 33% more tall ones (> 100 cm in height), and 66% more with smaller diameters (up to 50 cm) than beginners.
Although experience nearly doubled detection probability, experts were still imperfect observers. Detection probability may have been reduced because the survey was done early in the season during bouts of heavy rain. Dense undergrowth and steep topography (especially in plots 6, 12, and 14) also could have contributed to a high level of omissions. Nevertheless, detection probabilities of RWA nests in our study (Table 1) were comparable to those estimated in other studies of ants (Dorazio et al. 2011, Ward and Stanley 2013). Although standard surveys of RWA are done during the summer, the dense undergrowth present then could lead to a higher percentage of non-detection. In contrast, we suggest that sampling RWA nests would be better done in early spring when vegetation has not yet started to obscure the nests but temperatures are sufficiently high for ant activity. Finally, we found that with more surveys (or replicated ones: Dorazio et al. 2011), the combined detection probability increased relative to detection probability estimated from a single observer. However, the gain in detection probability of RWA nests showed diminishing returns beyond 6 – 8 observers (Fig. 7).

Even things as conspicuous as ant nests can be overlooked easily. Robust estimation of population density of sessile organisms—even highly apparent ones such as RWA nests—requires unbiased estimation of detection probability, just as it does when estimating population density of rare or cryptic species. Our Bayesian model for detection probability of sessile organisms included overlooked nests and other sources of heterogeneity in both occurrence and detection probabilities, and contributes to the further development of new methods for accurate assessments of population sizes.

As myrmecologists, we naturally are always surprised that not everyone is interested in mapping ant nests or estimating changes in ant population sizes through time and space (see also Underwood and Fisher 2006). However, the approach outlined here is relevant to any sessile
organism for which robust population estimates are desired but resources for exhaustive, repeated, population counts or estimates are limited (e.g., Philippi et al. 2001, Godefroid et al. 2011). Our methods can be used to provide answers to questions such as “how many surveyors do I need to accurately estimate the size of this population?” or “can I use non-expert surveyors, and how does that affect detection probability and their estimates of population size?” Answers to these questions are vital not only when censuses are conducted by experts, but also when citizen-scientists are engaged in mapping and monitoring of both common and rare species (e.g., Godet et al. 2009, Dickinson et al. 2010).

**Acknowledgements**

We thank our colleagues from Brasov-Romania University and Forest Research Station, and Klaus Bernd Nickel, chairman of Ameisenschutzwarte Hessen e.V., for their efforts and support during the field study. Gita Benadi worked out the probability of overlooking nests shown in Fig.7. Support for this work was provided by NSF grants 11-36646 and 12-37491 to AME and 11-36703 to NJS. NJS also thanks the Danish National Research Foundation for support provided to the Center for Macroecology, Evolution and Climate.

**Endangered species**

We certify that our research was conducted in conformance with all applicable laws.

**Literature Cited**


Berberich et al., 2016 – Detection probabilities for sessile organisms – 19

Insectes Sociaux 60:103–110.


Berberich et al., 2016 – Detection probabilities for sessile organisms – 21


R Core Team. 2015. R version 3.2.2. Available online: https://www.r-project.org/.


Berberich et al., 2016 – Detection probabilities for sessile organisms – 24

Tables

Table 1. Estimated detection probability $\hat{P}_i$ and its standard deviation (SD) for each of the eight observers (6 “Beginners” and 2 “Experts”), using the patch-occupancy model per-observer observation, Bayesian site-level detection model and site-level maximum likelihood model

<table>
<thead>
<tr>
<th>Method</th>
<th>“Beginners”</th>
<th>“Experts”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observer</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Patch-occupancy</td>
<td>$\hat{P}_i$</td>
<td>0.37 0.38</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.040 0.038</td>
</tr>
<tr>
<td>Maximum</td>
<td>$\hat{p}_i$</td>
<td>0.35 0.35</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.062 0.060</td>
</tr>
<tr>
<td>Bayes</td>
<td>$\hat{p}_i$</td>
<td>0.31 0.32</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.039 0.039</td>
</tr>
</tbody>
</table>
Table 2. Estimated number of nests $\overline{N}_s$ (maximum likelihood) in each plot and its standard deviation (SD), assuming a detection probability equal to the mean of the $\widehat{P}_i = 0.42$ (maximum likelihood) or 0.39 (Bayes) from Table 1.

<table>
<thead>
<tr>
<th>Plot</th>
<th>Maximum likelihood</th>
<th>Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\overline{N}_s$</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>27.5</td>
<td>6.25</td>
</tr>
<tr>
<td>2</td>
<td>15.7</td>
<td>3.36</td>
</tr>
<tr>
<td>3</td>
<td>11.3</td>
<td>2.73</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>23.1</td>
<td>4.57</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>0.31</td>
</tr>
<tr>
<td>7</td>
<td>0.6</td>
<td>0.09</td>
</tr>
<tr>
<td>8</td>
<td>6.2</td>
<td>1.91</td>
</tr>
<tr>
<td>9</td>
<td>9.2</td>
<td>2.78</td>
</tr>
<tr>
<td>10</td>
<td>19.8</td>
<td>3.92</td>
</tr>
<tr>
<td>11</td>
<td>0.3</td>
<td>0.05</td>
</tr>
<tr>
<td>12</td>
<td>18.3</td>
<td>3.44</td>
</tr>
<tr>
<td>13</td>
<td>3.7</td>
<td>1.32</td>
</tr>
<tr>
<td>14</td>
<td>4.9</td>
<td>1.46</td>
</tr>
<tr>
<td>15</td>
<td>15.3</td>
<td>3.86</td>
</tr>
<tr>
<td>16</td>
<td>17.8</td>
<td>3.54</td>
</tr>
</tbody>
</table>

$\overline{N} = 168.2$ \hspace{1cm} $190.1$
Table 3. Total number of nests ($\hat{N}$) and standard errors (SE) estimated over the 16 plots based on observations by 8 observers. Chao, Jackknife, and Bootstrap estimates were computed using the *specpool* function in the *vegan* package of R, version 3.2. Maximum likelihood and Bayesian estimates are from the column sums of Table 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{N}$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>Chao</td>
<td>163.7</td>
<td>8.25</td>
</tr>
<tr>
<td>Jackknife1</td>
<td>172.4</td>
<td>11.88</td>
</tr>
<tr>
<td>Jackknife2</td>
<td>180.0</td>
<td></td>
</tr>
<tr>
<td>Bootstrap</td>
<td>160.0</td>
<td>8.63</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>168.2</td>
<td></td>
</tr>
<tr>
<td>Bayes</td>
<td>190.1</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Analysis of variance table of the effects of stand characteristics and their interactions on frequency of detecting a red wood ant nest. Forest type was coded as spruce or not spruce; Location was coded as forest interior, forest edge, or forest path.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>1</td>
<td>168.3</td>
<td>32.85</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Diameter</td>
<td>1</td>
<td>6.56</td>
<td>1.28</td>
<td>0.26</td>
</tr>
<tr>
<td>Forest type</td>
<td>1</td>
<td>10.9</td>
<td>2.12</td>
<td>0.15</td>
</tr>
<tr>
<td>Location</td>
<td>2</td>
<td>10.5</td>
<td>2.04</td>
<td>0.13</td>
</tr>
<tr>
<td>Height × Forest type</td>
<td>1</td>
<td>5.1</td>
<td>0.99</td>
<td>0.32</td>
</tr>
<tr>
<td>Height × Location</td>
<td>2</td>
<td>1.4</td>
<td>0.27</td>
<td>0.76</td>
</tr>
<tr>
<td>Forest type × Location</td>
<td>1</td>
<td>2.7</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>Residuals</td>
<td>129</td>
<td>5.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure legends**

Fig.1. Photographs of RWA nest A_03_06 by each observer; different observers are indicated by their initials.

Fig. 2. Individual search tracks (colors denote different observers) recorded by GPS in the Fig. already scaled to max. width (180 mm)16 sampled plots. Letters denote roads (a); floodplain/wetland (b); dense understory layer of natural tree regeneration (c); dense understory of herbs and shrubs (d) and meeting point (e).

Fig. 3. Matrix of individual nests (rows) found (black) or undetected (white) by each observer (columns); nests are pooled across all 16 plots. Horizontal orange lines delimit nests seen (bottom to top) by all 8, only 7, …, 1 of the observers. White cells at the top represent nests undetected by all eight observers, based on the total number of nests across the 16 plots observed at least once ($S_{obs} = 147$) or estimated by Chao, jackknife 1 and 2, nest-level Bayesian patch-occupancy data-augmentation (Patch occ.), maximum likelihood (MLE) and Bayesian models (triangles). Colored circles at the left of each column represent the four nest height classes (1 – 10 (red), 11 – 50 (green), 51 – 100 (blue), and >101 cm (brown)).
Fig. 4. Proportion of successful detections as a function of nest height (left) or diameter (right). Point sizes are proportional to the number of nests (grey circles). Solid lines are best-fit logistic regressions (quasi-binomial general linear model); dotted lines are ± 2 standard errors. Note that nests in the largest size classes (height $\geq 100$ cm; diameter $\geq 200$ cm) are pooled in these figures.

Fig. 5. Effect of nest abundance at each plot on the probability of detection. No trend was detectable in these data.

Fig. 6. Stand characteristics and detectability of nests: Locations and numbers of plots and example of nest position and frequency of detection in Plot 1 (aerial photograph: Google earth).

Fig. 7. Maximum likelihood estimates of the number of nests at each plot, based on 100 randomly drawn combinations of 2 to 7 observers. Plots are sorted by number of estimated nests based on eight observers (•). A + indicates confirmed number of nests, Solid orange and red circles are, respectively, the estimated number of nests according to Chao and Jackknife1 estimators (Table 3). Inset: Simulated probability of overlooking a nest as a function of the number of observers. This simulation uses the data from the 16 plots, each bootstrapped 1000 times to simulate a random sequence of observers. Horizontal dashed lines are at 10%, 5%, and 1% of overlooked nests.
Fig. 1. Photographs of RWA nest A_03_06 by each observer; different observers are indicated by their initials.

156x69mm (300 x 300 DPI)
Fig. 2. Individual search tracks (colors denote different observers) recorded by GPS in the 16 sampled plots. Letters denote roads (a); floodplain/wetland (b); dense understory layer of natural tree regeneration (c); dense understory of herbs and shrubs (d) and meeting point (e).

156×140mm (300 x 300 DPI)
Fig. 3. Matrix of individual nests (rows) found (black) or undetected (white) by each observer (columns); nests are pooled across all 16 plots. Horizontal orange lines delimit nests seen (bottom to top) by all 8, only 7, ..., 1 of the observers. White cells at the top represent nests undetected by all eight observers, based on the total number of nests across the 16 plots observed at least once ($S_{obs} = 147$) or estimated by Chao, jackknife 1 and 2, nest-level Bayesian patch-occupancy data-augmentation (Patch occ.), maximum likelihood (MLE) and Bayesian models (triangles). Colored circles at the left of each column represent the four nest height classes (1 – 10 (red), 11 – 50 (green), 51 – 100 (blue), and >101 cm (brown)).
Fig. 4. Proportion of successful detections as a function of nest height (left) or diameter (right). Point sizes are proportional to the number of nests (grey circles). Solid lines are best-fit logistic regressions (quasi-binomial general linear model); dotted lines are ± 2 standard errors. Note that nests in the largest size classes (height ≥ 100 cm; diameter ≥ 200 cm) are pooled in these figures.
Fig. 5. Effect of nest abundance at each plot on the probability of detection. No trend was detectable in these data.
Fig. 6. Stand characteristics and detectability of nests: Locations and numbers of plots and example of nest position and frequency of detection in Plot 1 (aerial photograph: Google earth).

156x76mm (300 x 300 DPI)
Fig. 7. Maximum likelihood estimates of the number of nests at each plot, based on 100 randomly drawn combinations of 2 to 7 observers. Plots are sorted by number of estimated nests based on eight observers (●). A + indicates confirmed number of nests. Solid orange and red circles are, respectively, the estimated number of nests according to Chao and Jackknife1 estimators (Table 3). Inset: Simulated probability of overlooking a nest as a function of the number of observers. This simulation uses the data from the 16 plots, each bootstrapped 1000 times to simulate a random sequence of observers. Horizontal dashed lines are at 10%, 5%, and 1% of overlooked nests.
Appendix S1 – Supplement accompanying:
Detection probabilities for sessile organisms

G.M Berberich, C.F. Dormann, D. Klimetzek, M.B. Berberich, N.J. Sanders & A.M. Ellison

Detailed methods of simulating how many observers would be needed to estimate accurately the number of nests in a site

We started with the assumption that each observer had a probability $P_d$ of detecting a given RWA. Because we observed that some nests consistently were detected (or overlooked), we defined $P_c$ to be the probability that observer $i$ detected a RWA nest that was not detected by observer $j$, $i > j$. $P_c$ is “complementarity for zeros”, i.e., it is a conditional probability of finding a nest where the previous observer did not: $P_c = P(i+1 = 1 | i = 0)$. For a series of $n$ observers $i = \{i_1, i_2, \ldots, i_n\}$ visiting the same site $s$, the probability that a given nest has been overlooked is determined recursively:

\[
P(i_1 = 0) = 1 - P_d;
\]
\[
P(i_2, i_1 = 0) = P(i_2 = 0 | i_1 = 0) P(i_1 = 0) = (1-P_c) (1-P_d)
\]
\[
P(i_3 = 0, i_2 = 0, i_1 = 0) = P(i_3 = 0 | i_2 = 0, i_1 = 0) = (1-P_c) (1-P_c) (1-P_d) \quad \text{(B1)}
\]
\[
\ldots
\]
\[
P(i_n = 0, \ldots, i_1 = 0) = (1-P_c)^{n-1} (1-P_d).
\]

Intuitively, equation B1 means that the probability of $n$ observers overlooking a nest is dependent on the detection probability of the first, and the complementarity score of all
subsequent observers. We had an estimate of the (average) detection probability from our initial maximum-likelihood computations ($P_d = 0.42$; Table 1), so we computed, for any pair of observers, the proportion of visits the second observer found a nest that was previously undetected. On average, this quantity is $P_c = P(i_2 = 1 | i_1 = 0) = 0.25$. In other words, 65% of the effort of each additional observer could be considered to be redundant (wasted). The quantity $P_c$ quantified the correlation between observers and could not be expressed in terms of $P_d$. As both $P_d$ and $P_c$ were estimated from the data and hence were random variables, we bootstrapped the above function using random draws from the observed values of $P_d$ and $P_c$ to compute the variance. To compute $P_d$, we counted, for any pair of observers, how often a 0 of observer A was complemented by a 1 of observer B. Finally, we noted that as the number of observers, $n$, increased, the probability of overlooking any individual nest decreased.

With these estimates in hand, we then asked: how many observers would be needed to come within $x$ nests of the true number of nests, $N_s$, i.e., to reduce the probability of overlooking a nest $P(i_n = 0, \ldots i_1 = 0) = (1-P_c)^n-1(1-P_d)$ to less than a fixed quantity (e.g., 10%). The inset in Fig. 3 (main text) shows the bootstrapped probability of overlooking a nest with indicated targets at 10%, 5%, and 1% (horizontal dashed lines in the inset to Fig. 3 in the main text). To determine these values, we randomly drew an observer (with a detection probability determined from Table 1), drew a second observer randomly, looked up the overlooking rate for the second observer given the first observer (computed from the data), then drew another and so on. This simulation was repeated 1000 times (R code provided in the Supplement).
1 Introduction

This document contains the statistical analyses accompanying the paper “Detection probabilities for sessile organisms” by Berberich et al (2016). It presents the R-code and results for full reproducibility.

The analysis is carried out in three parts:

1. Evaluation whether nest size or plot-level predictors have an effect on detection.
2. The analysis of the number of nests across all plots using patch-occupancy models with data augmentation and a event-specific covariate (nest size); this results in estimates of how many nests were overlooked in total.
3. A Bayesian detection probability model across observers, based on a binomial sampling model. This model does not include nest sizes and is thus much simpler to implement.
4. A maximum likelihood version of the previous model. We use this model to quickly run the analysis for different sets of observers. It would take years to run model 1 for thousands of combinations of observers, and hence we had to resort to a maximum likelihood version. In fact, model 2 primarily serves as a link between these two models, illustrating that the maximum-likelihood model yields estimates similar to model 2, and that the main benefit of the data-augmentation approach is the incorporation of nest sizes.

2 The data

We have three data sets: plots, nests, and sizes. plots contains the misidentification-corrected recorded nests \( \left( N_{\text{obs}}^{i,s} \right) \) for each of the 8 observers (in columns: O1 to O8) for each of the 16 plots (in rows: Plot 1 to Plot 16). As additional columns it contains the total number of different nests recorded at each plot, which is our lower bound \( \left( N_{\text{obs}}^{s} \right) \) for the true number of nests at each plot \( \left( \hat{N}_{s} \right) \).

```r
plots <- read.csv("nestPlots.csv", row.names = 1)
```

The second data set, nests, is a long version of plots, in that it contains for each observer the information which nests he/she has detected (actual confirmed nests only).

```r
nests <- read.csv("nestRecords.csv", row.names = 1)
```

The third data set contains the nest sizes estimated roughly from the photographs (height, in cm, along with the variables diameters, locations and forest setting).

```r
sizes <- read.csv("nestSizes.csv", row.names = 1)  # read file in again to get all nest sizes
sizes$Height[which(sizes$Height > 100)] <- 100  # moves 1 nest to smaller size
sizes$Diameter[which(sizes$Diameter > 200)] <- 200  # moves 3 nests

# read file in again to get all nest sizes

# moves 3 nests

```

3 Effect of covariates on detection probability

3.1 Univariate exploration of predictors for detection probability

Across all observers, nest size (height or diameter) or landscape setting (location, forest type) may affect detection. Here we use a GLM to find out. First, for each nest we compute how many observers detected it. Then we relate this proportion to nest size, etc.

3.1.1 Nest size and diameter

```r
# join tables (sorted in the same way):
detnetsize <- cbind.data.frame(rowSums(nests), 8 - rowSums(nests), sizes)
summary(fmHeight <- glm(as.matrix(detnetsize[, 1:2]) ~ poly(Height, 2), family = quasibinomial, data = detnetsize))

Call:
glm(formula = as.matrix(detnetsize[, 1:2]) ~ poly(Height, 2),
```
family = quasibinomial, data = detnetsize)

Deviance Residuals:
  Min 1Q Median 3Q Max
-3.1474 -1.4109 -0.1274 1.4317 4.1385

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.01676 | 0.10023 | 0.167 | 0.86744 |
| poly(Height, 2)1 | 6.73074 | 1.21899 | 5.522 | 1.52e-07 *** |
| poly(Height, 2)2 | -3.77773 | 1.21330 | -3.114 | 0.00223 ** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 2.673435)

Null deviance: 556.86 on 146 degrees of freedom
Residual deviance: 443.53 on 144 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

predsHeight <- predict(fmHeight, newdata = data.frame(Height = 10:100), se.fit = T)

summary(fmDiameter <- glm(as.matrix(detnetsize[, 1:2]) ~ poly(Diameter, 2), family = quasibinomial, data = detnetsize))

Call:
  glm(formula = as.matrix(detnetsize[, 1:2]) ~ poly(Diameter, 2),
      family = quasibinomial, data = detnetsize)

Deviance Residuals:
  Min 1Q Median 3Q Max
-3.5483 -1.1682 -0.1953 1.4989 3.7833

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.03257 | 0.10276 | 0.317 | 0.752 |
| poly(Diameter, 2)1 | 6.66866 | 1.37642 | 4.845 | 3.24e-06 *** |
| poly(Diameter, 2)2 | -0.26763 | 1.36859 | -0.196 | 0.845 |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 2.865521)

Null deviance: 556.86 on 146 degrees of freedom
Residual deviance: 477.70 on 144 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

3
```r
predsDiameter <- predict(fmDiameter, newdata = data.frame(Diameter = 50:200), 
                        se.fit = T)
with(detnetsize, cor(Height, Diameter))

[1] 0.7910949

We can use a bubble plot to visualise this.

```
Symbol size is proportional to the number of nests of that combination of size and numbers of observers that discovered it.

Since diameter and height are highly correlated, and size is the better predictor, we shall henceforth only use height to represent size.

### 3.1.2 Location and forest type

All but three nests were recorded in spruce forest (one in pine, one in beech), and hence we would not expect to be able to detect effects of forest type. Similarly, location has several levels (13), but 95/147 data points are from fully surrounded by forest, rather than moss, thistle etc.

```r
anova(glm(as.matrix(detnetsize[, 1:2]) ~ Forest, family = quasibinomial, data = detnetsize), test = "F")
```

Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: as.matrix(detnetsize[, 1:2])

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>138</td>
<td>530.50</td>
<td>530.50</td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>Forest</td>
<td>2</td>
<td>7.509</td>
<td>136</td>
<td>522.99 1.1827 0.3096</td>
<td></td>
</tr>
</tbody>
</table>
### Nest size into detection rate analysis

For each nest, the probability of observing it depends on (a) the detection rate of the observer, (b) the size of the nest, and (c) plot characteristics. As shown in the last section, we have not recorded any useful measures of plot characteristics, so we leave out point (c) here.

We try two different models: `fmm1` with an observer-specific detection curve, and `fmm2` with the same detection curve for all observers, but an observer-specific intercept. The latter model will use fewer degrees of freedom. As this turns out to be the more appropriate model for our data, we plot these results.

```r
# reformat data for analysis: all observers underneath each other:
part1 <- stack(nests)
colnames(part1) <- c("detected", "observer")
part2 <- do.call("rbind", replicate(8, sizes, simplify = FALSE))
dats <- cbind(part1, part2)
# head(dats)
library(lme4)
# fit a model with variable effect of nest height for each
# observer:
summary(fmm1 <- glmer(detected ~ (poly(Height, 2) | observer), family = binomial, 
                            data = dats))
```

Generalized linear mixed model fit by maximum likelihood
  (Laplace Approximation) [glmerMod]
Family: binomial  (logit)
Formula: detected ~ (poly(Height, 2) | observer)
  Data: dats

AIC BIC logLik deviance df.resid
1525.7 1561.2  -755.8  1511.7   1169

Scaled residuals:
  Min   1Q Median   3Q  Max
-1.6253 -0.7663  0.6153  0.7934  1.8713
Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>observer</td>
<td>(Intercept)</td>
<td>1.212</td>
<td>1.101</td>
<td></td>
</tr>
<tr>
<td></td>
<td>poly(Height, 2)1</td>
<td>379.454</td>
<td>19.480</td>
<td>-0.99</td>
</tr>
<tr>
<td></td>
<td>poly(Height, 2)2</td>
<td>129.448</td>
<td>11.378</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Number of obs: 1176, groups: observer, 8

Fixed effects:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 1.028 | 0.480   | 2.14     | 0.0323 * |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```r
glmer(detected ~ poly(Height, 2) + (1 | observer),
      family = binomial, data = dats)
```

Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]

Family: binomial (logit)
Formula: detected ~ poly(Height, 2) + (1 | observer)
Data: dats

AIC BIC logLik deviance df.resid
1498.2 1518.5 -745.1 1490.2 1172

Scaled residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.9525</td>
<td>-0.7399</td>
<td>0.5122</td>
<td>0.8212</td>
<td>1.7025</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>observer</td>
<td>(Intercept)</td>
<td>0.1622</td>
<td>0.4028</td>
</tr>
</tbody>
</table>

Number of obs: 1176, groups: observer, 8

Fixed effects:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 0.01972 | 0.15548  | 0.127   | 0.899 |
| poly(Height, 2)1 | 19.76902 | 2.17022  | 9.109 < 2e-16 *** |
| poly(Height, 2)2 | -11.09957| 2.14672  | -5.170 2.34e-07 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th>(Intr)</th>
<th>p(H,2)1</th>
<th>ply(Hgh,2)1</th>
<th>ply(Hgh,2)2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>-0.004</td>
</tr>
</tbody>
</table>

```r
plot(10:100, plogis(predict(fmm2, newdata = data.frame(Height = 10:100, observer = "O1"))), type = "l", las = 1, xlab = c("10", 100), ylim = c(-5.170, 2.34e-07 ***), ylab = "detection rate", xlab = "nest height [cm]", axes = F)
```

axis(1, at = c(10, 50, 100), labels = c("10", "50", "100+"))
axis(2, las = 1)
All the analyses above are fine for analysing the correlation between detection of nests and various attributes, but they do not tell us anything about how many nests we have not seen. To answer that question, we turn to a very different approach. The above analyses have been useful, however, in guiding us which covariates to include in the following step.

4 Data augmented patch-occupancy model with event-specific covariate

Our data provide the following challenges:

1. We have eight observers sampling the same plots, but each has a different detection rate (due to experience, eye sight, ...).
2. We have shown that small ant nests are easier to overlook than large ones. Thus, each event (“ant nest”) has a covariate affecting its detection (nest size, which we simplify to the values “small”=0 and “large”=1, for 10 cm and others, respectively).
3. We may have some nests that none of our eight observers discovered. For those we obviously also do not know the size.

Typical patch-occupancy data assume constant detection rates (“repeated within-season visits”) and focus on detection and occurrence of the (typically) animal at each plot. Instead, we want to estimate how many nests...
were not recorded at all. As ant nests (similar to trees, but in contrast to animals) don’t move, we can safely assume that occurrence (psi) is 1 if any observer has observed a nest.

We can handle the “overlooked nests”-issue by adding NA-records to our data set nests, which are then guessed (estimated) during the modelling procedure. This is called “data-augmentation”, which feels a bit like Bayesian magic, but isn’t. What the model does is to estimate for \( N^{aug} \) nests which were not observed, how likely it is that they are there, but were not observed. This can be achieved by realising that also the unobserved nests (and their sizes) are drawn from the same data model that we fit to the observed data. The main tuning parameter on top of a simpler patch-occupancy model is the number of nests we assume to be missing. (In the specific case, we shall assume \( N^{aug} = 50 \) overlooked nests, but the results do not change if we assume 20 or 200 instead.)

```r
library(R2jags) # load access to JAGS
# augment the matrix with some unobserved nests:
Nunobseved <- 50
augnests <- rbind(as.matrix(nests), matrix(0, Nunobseved, 8))
jags.data <- list(Y = augnests, N = NROW(augnests), J = NCOL(augnests),
                   nestsize = c(ifelse(sizes$Height < 70, 0, 1), rep(NA, Nunobseved))) # categorise nest size into small

augAnalysis <- function() {
  # the classical patch-occupancy model: loop through nests,
  # observed plus augmented
  for (i in 1:N) {
    w[i] ~ dbern(omega) # realised nest probability
    nestsize[i] ~ dbern(probnestsize) # either nest size 0 (small) or 1 (large)
    for (j in 1:J) {
      # loop through observers
      Y[i, j] ~ dbern(P[i, j] * w[i]) # compute detection based on the members in the set and the guesses for the unobserved nests
      logit(P[i, j]) <- detectrate[j] + betasize * nestsize[i] # nestsize effect on detection
    }
  }
  # Priors and constraints:
  for (j in 1:J) {
    detectrate[j] ~ dnorm(0, 0.01) # flat but informative prior centred on p=0.5
    # (note: this is at logit-scale, thus mu=0 -> p=0.5);
    # curve(plogis(dnorm(x, 0, 10)), -20, 20)
  }
  omega ~ dunif(0, 1)
  probnestsize ~ dbeta(1, 1)
  betasize ~ dnorm(0, 0.01)
  # derived parameters:
  Ntrulythere <- sum(w) # number of nests across all plots
  for (j in 1:J) {
    # back-transformed detection rate per observer
    detectionRateRealScale[j] <- exp(detectrate[j])/(1 + exp(detectrate[j]))
  }
}

# end of function

# inits<-function() list (w=c(rep(1, NROW(inventedData)), rep(0,
# Nunobserved), betasize=rnorm(1),
# detectrate=rnorm(n=NCOL(inventedAugnests),1))

parms <- c("omega", "Ntrulythere", "detectionRateRealScale", "betasize",
           "probnestsize")
ni <- 2000
nb <- ni/2
nc <- 3
nt <- 3  # 8000 will do for final estimation!

inits <- function() list(w = c(rep(1, NROW(nests)), rep(0, Nunobserved)),
                         betasize = rnorm(1), detectrate = rnorm(n = NCOL(augnests), 1))

# call JAGS
system.time(augJags <- jags(jags.data, inits, parms, model.file = augAnalysis,
                          n.chains = nc, n.thin = nt, n.iter = ni, n.burnin = nb, working.directory = getwd()))

module glm loaded

Compiling model graph
  Resolving undeclared variables
  Allocating nodes
Graph information:
  Observed stochastic nodes: 1723
  Unobserved stochastic nodes: 258
  Total graph size: 4507

Initializing model

  user  system elapsed
   10.979   0.066  11.161

plot(augJags)
### Inference for Bugs model at "/var/folders/cc/3jfhfx190rb2ptxnqrqxj94m0000gp/T//RtmpgqyR2o/model297244f75d69.txt", fit using jags, 3 chains, each with 2000 iterations (first 1000 discarded), n.thin = 3

n.sims = 1002 iterations saved

<table>
<thead>
<tr>
<th></th>
<th>mu.vect</th>
<th>sd.vect</th>
<th>2.5%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntruelythere</td>
<td>147.627</td>
<td>0.811</td>
<td>147.000</td>
<td>147.000</td>
</tr>
<tr>
<td>betasize</td>
<td>0.817</td>
<td>0.165</td>
<td>0.493</td>
<td>0.708</td>
</tr>
<tr>
<td>detectionRateRealScale[1]</td>
<td>0.366</td>
<td>0.042</td>
<td>0.285</td>
<td>0.339</td>
</tr>
<tr>
<td>detectionRateRealScale[2]</td>
<td>0.381</td>
<td>0.039</td>
<td>0.314</td>
<td>0.353</td>
</tr>
<tr>
<td>detectionRateRealScale[3]</td>
<td>0.437</td>
<td>0.041</td>
<td>0.352</td>
<td>0.411</td>
</tr>
<tr>
<td>detectionRateRealScale[4]</td>
<td>0.407</td>
<td>0.040</td>
<td>0.332</td>
<td>0.380</td>
</tr>
<tr>
<td>detectionRateRealScale[5]</td>
<td>0.488</td>
<td>0.044</td>
<td>0.401</td>
<td>0.457</td>
</tr>
<tr>
<td>detectionRateRealScale[6]</td>
<td>0.426</td>
<td>0.041</td>
<td>0.347</td>
<td>0.397</td>
</tr>
<tr>
<td>detectionRateRealScale[7]</td>
<td>0.613</td>
<td>0.041</td>
<td>0.530</td>
<td>0.585</td>
</tr>
<tr>
<td>detectionRateRealScale[8]</td>
<td>0.644</td>
<td>0.040</td>
<td>0.565</td>
<td>0.617</td>
</tr>
<tr>
<td>omega</td>
<td>0.748</td>
<td>0.031</td>
<td>0.679</td>
<td>0.727</td>
</tr>
<tr>
<td>probnestsize</td>
<td>0.175</td>
<td>0.031</td>
<td>0.119</td>
<td>0.155</td>
</tr>
<tr>
<td>deviance</td>
<td>1711.737</td>
<td>9.690</td>
<td>1699.031</td>
<td>1704.103</td>
</tr>
</tbody>
</table>

50% 75% 97.5% Rhat n.eff

<table>
<thead>
<tr>
<th></th>
<th>Ntruelythere</th>
<th>betasize</th>
<th>detectionRateRealScale[1]</th>
<th>detectionRateRealScale[2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>147.000</td>
<td>148.000</td>
<td>0.365</td>
<td>0.381</td>
</tr>
<tr>
<td>75%</td>
<td>150.000</td>
<td>1.131</td>
<td>0.446</td>
<td>1.001</td>
</tr>
<tr>
<td>97.5%</td>
<td>1.004</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>
For each parameter, n.eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 46.9 and DIC = 1758.6
DIC is an estimate of expected predictive error (lower deviance is better).

The results show that large nests have a higher chance of being detected (notice the estimate for betasize of 0.819 at the link scale, representing the effect of going from small to large nests). Furthermore, we get per-observer observation estimates (at the real scale) between 0.365 and 0.642, i.e. almost a factor of 2. And, finally, we get an estimate of the total number of nests across all plots as 147.7 (95%-confidence interval up to 150), i.e. 1 to 3 nests overlooked.

Since we do not have any covariates at the plot level, we can distribute the overlooked nests across plots proportional to the number of nests observed there.

```r
detectionRateRealScale[3]  0.437  0.463  0.519  1.007  400
detectionRateRealScale[4]  0.406  0.434  0.489  1.000  1000
detectionRateRealScale[5]  0.489  0.517  0.574  1.000  1000
detectionRateRealScale[6]  0.424  0.452  0.513  1.000  1000
detectionRateRealScale[7]  0.613  0.642  0.693  1.001  1000
detectionRateRealScale[8]  0.642  0.672  0.723  1.003  790
omega                    0.747  0.769  0.807  1.002  800
probnestsize             0.174  0.194  0.238  1.000  1000
deviance                  1710.662 1717.094 1736.772 1.002  710
```

```r
quants <- quantile(augJags$BUGSoutput$sims.list$Ntruelythere - 147, c(0.025, 0.5, 0.975))
# So the number of nests per plot are:
estimated <- matrix(plots$Nmin, ncol = 3, nrow = 16, byrow = F) +
              matrix(quants, ncol = 3, nrow = 16, byrow = T)/matrix(plots$Nmin,
              ncol = 3, nrow = 16, byrow = F)
# in line with reason, but against maths, we assume for plot 4
# that 0/0=0, and get:
estimated[4, ] <- 0
colnames(estimated) <- c("lower CI", "median", "upper CI")
round(estimated, 2)
```

<table>
<thead>
<tr>
<th></th>
<th>lower CI</th>
<th>median</th>
<th>upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>24</td>
<td>24</td>
<td>24.12</td>
</tr>
<tr>
<td>[2,]</td>
<td>17</td>
<td>17</td>
<td>17.18</td>
</tr>
<tr>
<td>[3,]</td>
<td>8</td>
<td>8</td>
<td>8.38</td>
</tr>
<tr>
<td>[4,]</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>[5,]</td>
<td>18</td>
<td>18</td>
<td>18.17</td>
</tr>
<tr>
<td>[6,]</td>
<td>3</td>
<td>3</td>
<td>4.00</td>
</tr>
<tr>
<td>[7,]</td>
<td>1</td>
<td>1</td>
<td>4.00</td>
</tr>
<tr>
<td>[8,]</td>
<td>4</td>
<td>4</td>
<td>4.75</td>
</tr>
<tr>
<td>[9,]</td>
<td>10</td>
<td>10</td>
<td>10.30</td>
</tr>
<tr>
<td>[10,]</td>
<td>20</td>
<td>20</td>
<td>20.15</td>
</tr>
<tr>
<td>[11,]</td>
<td>1</td>
<td>1</td>
<td>4.00</td>
</tr>
<tr>
<td>[12,]</td>
<td>11</td>
<td>11</td>
<td>11.27</td>
</tr>
<tr>
<td>[13,]</td>
<td>4</td>
<td>4</td>
<td>4.75</td>
</tr>
<tr>
<td>[14,]</td>
<td>3</td>
<td>3</td>
<td>4.00</td>
</tr>
<tr>
<td>[15,]</td>
<td>12</td>
<td>12</td>
<td>12.25</td>
</tr>
</tbody>
</table>
Essentially this indicates that with our eight observers, we have good faith of not having overlooked any nest!

One advantage of the above data-augmentation approach is that it allows us to model each nest separately and thereby include a covariate for the nest. If we omit the effect of nest size, the results are as follows:

```r
augAnalysis2 <- function() {
  # the classical patch-occupancy model: loop through nests, 
  # observed plus augmented
  for (i in 1:N) {
    w[i] ~ dbern(omega)  # realised nest probability
    for (j in 1:J) {
      # loop through observers
      Y[i, j] ~ dbern(P[i, j] * w[i])  # compute detection based on the members in the set and the guesses for the unobserved nests
      logit(P[i, j]) <- detectrate[j]  #+ betasize*nestsize[i] # no nestsize effect on detection
    }
  }
  # Priors and constraints:
  for (j in 1:J) {
    detectrate[j] ~ dnorm(0, 0.01)  # flat but informative prior centred on p=0.5
    # (note: this is at logit-scale, thus mu=0 -> p=0.5);
    # curve(plogis(dnorm(x, 0, 10)), -20, 20)
  }
  omega ~ dunif(0, 1)
  # derived parameters:
  Ntruelythere <- sum(w)  # number of nests across all plots
  for (j in 1:J) {
    # back-transformed detection rate per observer
    detectionRateRealScale[j] <- exp(detectrate[j])/(1 + exp(detectrate[j]))
  }
}
# end of function

# inits<-function() list (w=c(rep(1, NROW(inventedData)), rep(0, # Nunobserved)), betasize=rnorm(1),
# detectrate=rnorm(n=NCOL(inventedAugnests),1))

parms <- c("omega", "Ntruelythere", "detectionRateRealScale")
inits <- function() list(w = c(rep(1, NROW(nests)), rep(0, Nunobserved)),
                         detectrate = rnorm(n = NCOL(augnests), 1))

# call JAGS
system.time(augJags2 <- jags(jags.data, inits, parms, model.file = augAnalysis2,
                          n.chains = nc, n.thin = nt, n.iter = ni, n.burnin = nb, working.directory = getwd()))
```

Warning in jags.model(model.file, data = data, inits =
init.values, n.chains = n.chains, : Unused variable "nestsize" in data
Compiling model graph
  Resolving undeclared variables
  Allocating nodes
Graph information:
  Observed stochastic nodes: 1576
  Unobserved stochastic nodes: 206
  Total graph size: 3428

Initializing model

```
user  system elapsed
4.685  0.068  4.965
```

```
plot(augJags2)
```

Inference for Bugs model at "/var/folders/cc/3jfhx190rb2ptxnqrqj94m0000gp/T//RtmpgqyR2o/model297226e112d8.txt", fit using jags, 3 chains, each with 2000 iterations (first 1000 discarded), n.thin = 3
n.sims = 1002 iterations saved

<table>
<thead>
<tr>
<th>mu.vect</th>
<th>sd.vect</th>
<th>2.5%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntruelythere</td>
<td>147.460</td>
<td>0.726</td>
<td>147.000</td>
</tr>
<tr>
<td>detectionRateRealScale[1]</td>
<td>0.403</td>
<td>0.040</td>
<td>0.328</td>
</tr>
</tbody>
</table>
detectionRateRealScale[2] 0.413 0.041 0.334 0.386
detectionRateRealScale[3] 0.467 0.041 0.389 0.438
detectionRateRealScale[4] 0.436 0.042 0.354 0.408
detectionRateRealScale[5] 0.523 0.042 0.441 0.494
detectionRateRealScale[6] 0.461 0.041 0.387 0.435
detectionRateRealScale[7] 0.643 0.038 0.563 0.618
detectionRateRealScale[8] 0.672 0.038 0.596 0.647
omega 0.745 0.031 0.686 0.724
deviance 1599.104 9.295 1588.527 1592.232

50% 75% 97.5% Rhat n.eff
Ntruelythere 147.000 148.000 149.000 1.001 1000
detectionRateRealScale[1] 0.403 0.431 0.482 1.004 480
detectionRateRealScale[2] 0.413 0.441 0.491 1.003 540
detectionRateRealScale[3] 0.469 0.496 0.548 1.000 1000
detectionRateRealScale[4] 0.436 0.466 0.513 1.004 480
detectionRateRealScale[5] 0.523 0.551 0.603 1.003 630
detectionRateRealScale[6] 0.461 0.488 0.542 1.003 1000
detectionRateRealScale[7] 0.644 0.669 0.715 1.009 220
detectionRateRealScale[8] 0.674 0.698 0.743 1.013 170
omega 0.746 0.766 0.802 1.003 530
deviance 1596.090 1604.094 1622.298 1.002 1000

For each parameter, n.eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 43.2 and DIC = 1642.3
DIC is an estimate of expected predictive error (lower deviance is better).

That means that detection probabilities for the observers are slightly higher across the board, but the estimated number of nests hardly changes. In other words, while nest size affects detection probability, it does not greatly bias estimates of the total number of nests.

One disadvantage of the above analysis is that we essentially treat the data as if they were from one large plot, rather than from 16 different plots. If we wanted to estimate abundance of red wood ant nests at each plot, rather than across all plots, we have to employ an abundance model. The approach of Royle (2004: N-mixture models for estimating population size from spatially replicated counts. Biometrics 60:108) is in principle suitable for such data, except that in his case the observation matrix is very sparse (has many 0s), while ours is dense (0s only in few observer-plot combinations):

plots

<table>
<thead>
<tr>
<th>01 02 03 04 05 06 07 08 Nmin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot 1 12 14 12 9 10 11 12 13 24</td>
</tr>
<tr>
<td>Plot 2 1 5 5 7 7 9 12 17</td>
</tr>
<tr>
<td>Plot 3 4 5 2 5 4 5 7 6 8</td>
</tr>
<tr>
<td>Plot 4 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Plot 5 9 7 7 6 13 8 14 14 18</td>
</tr>
<tr>
<td>Plot 6 0 0 0 0 0 0 3 2 3</td>
</tr>
<tr>
<td>Plot 7 0 0 1 0 0 0 1 0 1</td>
</tr>
<tr>
<td>Plot 8 2 2 2 3 4 2 3 3 4</td>
</tr>
<tr>
<td>Plot 9 3 3 3 4 3 5 5 7 10</td>
</tr>
<tr>
<td>Plot 10 6 5 10 8 9 8 9 10 20</td>
</tr>
<tr>
<td>Plot 11 0 0 0 1 0 0 0 0 1</td>
</tr>
</tbody>
</table>
In this case it seems more intuitive to model abundance of nests per plot as a binomial random variate, rather than the more involved mixture of \( N \) Poisson distributions. This is what our next model does.

```r
library(unmarked)
obs <- matrix(as.character(1:8), 16, 8, byrow = T)
sitecovs <- data.frame(X = as.factor(1:16))
antsumf <- unmarkedFramePCount(y = as.matrix(plots[, -9]), obsCovs = list(observer = obs), siteCovs = sitecovs)
(fit <- pcount(~observer ~ X, data = antsumf, K = 50, mixture = "P"))
```

Call:
```
pcount(formula = ~observer ~ X, data = antsumf, K = 50, mixture = "P")
```

Abundance:
```
  Estimate    SE      z   P(>|z|)  
(Intercept)  3.760 0.269 13.9927 1.73e-44  
    X2 -0.530 0.289 -1.8360 6.64e-02  
    X3 -0.889 0.327 -2.7162 6.60e-03  
    X4 -15.401 351.669 -0.0438 9.65e-01  
    X5 -0.156 0.259 -0.6035 5.46e-01  
    X6 -2.497 0.607 -4.1172 3.83e-05  
    X7 -3.589 1.015 -3.5378 4.03e-04  
    X8 -1.500 0.415 -3.6128 3.03e-04  
    X9 -1.027 0.344 -2.9879 2.81e-03  
   X10 -0.352 0.275 -1.2818 2.00e-01  
   X11 -3.674 1.025 -3.5846 3.38e-04  
   X12 -0.402 0.279 -1.4398 1.50e-01  
   X13 -1.931 0.494 -3.9078 9.31e-05  
   X14 -1.719 0.456 -3.7726 1.21e-04  
   X15 -0.569 0.293 -1.9400 5.24e-02  
   X16 -0.438 0.282 -1.5501 1.21e-01  
```

Detection:
```
  Estimate    SE      z   P(>|z|)  
(Intercept) -1.3132 0.326 -4.034 5.49e-05  
  observer2  0.0425 0.206  0.206 8.37e-01  
  observer3  0.2033 0.202  1.004 3.15e-01  
  observer4  0.1246 0.204  0.611 5.41e-01  
  observer5  0.3519 0.200  1.757 7.90e-02  
  observer6  0.1839 0.203  0.907 3.64e-01  
  observer7  0.6556 0.202  3.248 1.16e-03  
  observer8  0.7189 0.203  3.534 4.10e-04  
```

AIC: 449.1964
ests <- fit@estimates$estimates$det@estimates
# observer probabilities:
plogis(c(observer1 = unname(ests[1]), ests[2:8] + ests[1]))

observer1 observer2 observer3 observer4 observer5 observer6
0.2119490 0.2191383 0.2478838 0.2335077 0.2766223 0.2442902
observer7 observer8
0.3412760 0.3556458

# site estimates:
estplot <- fit@estimates$estimates$state@estimates
round(exp(c(X1 = unname(estplot[1]), estplot[2:16] + estplot[1])),
       2)

X1   X2   X3   X4   X5   X6   X7   X8   X9  X10
42.97 25.29 17.66  0.00  3.54  1.19  9.58 15.38 30.20
X11  X12  X13  X14  X15  X16
1.09 28.75  6.23  7.70 24.32 27.73

5 Bayesian plot-level detection model

Our model of the plots data as displayed above assumes that the number of nests observed in a plot $s$ by observer $i$ is a draw from a binomial distribution with an estimated population size $\hat{N}_{i,s}^{true}$ for each plot $s$, and an estimated observation probability $\hat{P}_i$ for observer $i$. Since we have eight observers and 16 plots, we can estimate both $\hat{P}_i$ and $\hat{N}_{i,s}$. For this Bayesian model we need to choose priors for detection probabilities and $\hat{N}_{i,s}$. The latter has a lower bound at $N_{i,s}^{min} = N_{i,s}^{obs}$, as there cannot be fewer nests than observed.

detectBinom <- function() {
  # the detection model loop through observers loop through plots
  for (i in 1:8) {
    for (j in 1:16) {
      Nobs[j, i] ~ dbin(Pi[i], Ntrue[j])
    }
  }

  # Priors and constraints:
  for (j in 1:16) {
    # Ntrue must be an integer greater or equal Nobs:
    Ntrue[j] ~ dpois(Nmin[j] * (1 + Propoverlooked[j]))
    T(Nmin[j], )
    # overlooked nests modelled as proportion of the number observed:
    Propoverlooked[j] ~ dexp(shapeOverlooked)
  }
  shapeOverlooked ~ dgamma(1, 1)
  # uninformative prior on detection:
  for (i in 1:8) {
    Pi[i] ~ dbeta(1, 1)  # flat line
  }

  # compute another value of interest:
  meanPropOverlooked <- mean(Propoverlooked)
}

# end of function
jags.data <- list(Nobs = plots[, -9], Nmin = plots[, 9])
parametersBinom <- c("Ntrue", "Pi", "Propoverlooked", "shapeOverlooked",
                   "meanPropOverlooked")

ni <- 8000
nb <- ni/2
nc <- 3
nt <- 3

# call JAGS
system.time(antdetectBinom <- jags(jags.data, inits = NULL, parametersBinom,
                                   model.file = detectBinom, n.chains = nc, n.thin = nt, n.iter = ni,
                                   n.burnin = nb, working.directory = getwd()))

plot(antdetectBinom)
Inference for Bugs model at "/var/folders/cc/3jhfx190rb2pxnqrqy94m0000gp/T//RtmpqygR2q/model29725dd4e1e5.txt", fit using jags, 3 chains, each with 8000 iterations (first 4000 discarded), n.thin = 3
n.sims = 4002 iterations saved

$\begin{array}{llllll}
\text{mu.vect} & \text{sd.vect} & \text{2.5\%} & \text{25\%} & \text{50\%} \\
\text{Ntrue[1]} & 29.077 & 2.893 & 24.000 & 27.000 & 29.000 \\
\text{Ntrue[2]} & 18.482 & 1.492 & 17.000 & 17.000 & 18.000 \\
\text{Ntrue[3]} & 11.670 & 1.586 & 9.000 & 11.000 & 12.000 \\
\text{Ntrue[4]} & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
\text{Ntrue[5]} & 24.779 & 2.557 & 20.000 & 23.000 & 25.000 \\
\text{Ntrue[6]} & 3.133 & 0.360 & 3.000 & 3.000 & 3.000 \\
\end{array}$
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>2.5%</th>
<th>97.5%</th>
<th>Rhat</th>
<th>n.eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntrue[7]</td>
<td>1.044</td>
<td>0.206</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[8]</td>
<td>6.158</td>
<td>1.061</td>
<td>4.000</td>
<td>5.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[9]</td>
<td>11.124</td>
<td>1.164</td>
<td>10.000</td>
<td>10.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[10]</td>
<td>21.704</td>
<td>1.691</td>
<td>20.000</td>
<td>20.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[11]</td>
<td>1.023</td>
<td>0.156</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[12]</td>
<td>18.695</td>
<td>2.161</td>
<td>15.000</td>
<td>17.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[13]</td>
<td>4.572</td>
<td>0.741</td>
<td>3.000</td>
<td>4.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[14]</td>
<td>4.883</td>
<td>0.939</td>
<td>3.000</td>
<td>4.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[15]</td>
<td>16.254</td>
<td>1.951</td>
<td>13.000</td>
<td>15.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[16]</td>
<td>18.006</td>
<td>2.143</td>
<td>14.000</td>
<td>17.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pi[1]</td>
<td>0.313</td>
<td>0.038</td>
<td>0.241</td>
<td>0.287</td>
<td>0.311</td>
<td></td>
</tr>
<tr>
<td>Pi[2]</td>
<td>0.323</td>
<td>0.038</td>
<td>0.252</td>
<td>0.297</td>
<td>0.322</td>
<td></td>
</tr>
<tr>
<td>Pi[3]</td>
<td>0.364</td>
<td>0.041</td>
<td>0.287</td>
<td>0.337</td>
<td>0.364</td>
<td></td>
</tr>
<tr>
<td>Pi[4]</td>
<td>0.345</td>
<td>0.040</td>
<td>0.271</td>
<td>0.317</td>
<td>0.343</td>
<td></td>
</tr>
<tr>
<td>Pi[5]</td>
<td>0.406</td>
<td>0.042</td>
<td>0.322</td>
<td>0.377</td>
<td>0.406</td>
<td></td>
</tr>
<tr>
<td>Pi[6]</td>
<td>0.359</td>
<td>0.040</td>
<td>0.280</td>
<td>0.331</td>
<td>0.358</td>
<td></td>
</tr>
<tr>
<td>Pi[7]</td>
<td>0.499</td>
<td>0.047</td>
<td>0.407</td>
<td>0.467</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>Pi[8]</td>
<td>0.520</td>
<td>0.048</td>
<td>0.425</td>
<td>0.488</td>
<td>0.520</td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[1]</td>
<td>0.186</td>
<td>0.164</td>
<td>0.060</td>
<td>0.141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[2]</td>
<td>0.145</td>
<td>0.139</td>
<td>0.044</td>
<td>0.104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[3]</td>
<td>0.278</td>
<td>0.258</td>
<td>0.085</td>
<td>0.205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[4]</td>
<td>0.315</td>
<td>0.347</td>
<td>0.084</td>
<td>0.208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[5]</td>
<td>0.259</td>
<td>0.213</td>
<td>0.092</td>
<td>0.203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[6]</td>
<td>0.231</td>
<td>0.250</td>
<td>0.064</td>
<td>0.156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[7]</td>
<td>0.263</td>
<td>0.294</td>
<td>0.068</td>
<td>0.168</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[8]</td>
<td>0.282</td>
<td>0.276</td>
<td>0.085</td>
<td>0.198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[9]</td>
<td>0.181</td>
<td>0.174</td>
<td>0.053</td>
<td>0.127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[10]</td>
<td>0.144</td>
<td>0.138</td>
<td>0.043</td>
<td>0.105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[11]</td>
<td>0.257</td>
<td>0.274</td>
<td>0.067</td>
<td>0.167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[12]</td>
<td>0.410</td>
<td>0.329</td>
<td>0.157</td>
<td>0.338</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[13]</td>
<td>0.215</td>
<td>0.214</td>
<td>0.062</td>
<td>0.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[14]</td>
<td>0.285</td>
<td>0.277</td>
<td>0.082</td>
<td>0.198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[15]</td>
<td>0.255</td>
<td>0.228</td>
<td>0.081</td>
<td>0.192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propoverlooked[16]</td>
<td>0.360</td>
<td>0.302</td>
<td>0.124</td>
<td>0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td>meanPropOverlooked</td>
<td>0.254</td>
<td>0.097</td>
<td>0.111</td>
<td>0.184</td>
<td>0.239</td>
<td></td>
</tr>
<tr>
<td>shapeOverlooked</td>
<td>3.654</td>
<td>1.385</td>
<td>1.590</td>
<td>2.648</td>
<td>3.448</td>
<td></td>
</tr>
<tr>
<td>deviance</td>
<td>391.993</td>
<td>6.101</td>
<td>381.722</td>
<td>387.700</td>
<td>391.340</td>
<td></td>
</tr>
</tbody>
</table>

75% Rhat n.eff

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>2.5%</th>
<th>97.5%</th>
<th>Rhat</th>
<th>n.eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntrue[1]</td>
<td>31.000</td>
<td>36.000</td>
<td>1.002</td>
<td>1400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[2]</td>
<td>19.000</td>
<td>22.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[3]</td>
<td>13.000</td>
<td>15.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[4]</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[6]</td>
<td>3.000</td>
<td>4.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[7]</td>
<td>1.000</td>
<td>2.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[8]</td>
<td>7.000</td>
<td>8.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[9]</td>
<td>12.000</td>
<td>14.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[10]</td>
<td>23.000</td>
<td>26.000</td>
<td>1.003</td>
<td>980</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[11]</td>
<td>1.000</td>
<td>1.000</td>
<td>1.003</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[12]</td>
<td>20.000</td>
<td>23.000</td>
<td>1.002</td>
<td>1300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[13]</td>
<td>5.000</td>
<td>6.000</td>
<td>1.002</td>
<td>2100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[14]</td>
<td>5.000</td>
<td>7.000</td>
<td>1.001</td>
<td>2500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[15]</td>
<td>17.000</td>
<td>20.975</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ntrue[16]</td>
<td>19.000</td>
<td>23.000</td>
<td>1.001</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For each parameter, n.eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 18.6 and DIC = 410.6
DIC is an estimate of expected predictive error (lower deviance is better).

---

The key findings are that we overlook, according to this model, around 26% of nests (estimating the total as roughly 191 ± 18); and that the detection rates vary between 0.312 and 0.520 among observers. These values look rather different to the previous data-augmentation model. This difference can be attributed to two changes in the model structure:

1. The current model has fewer data points because it aggregates all nests within a plot (16 · 8 = 128 vs. 147 for the data augmentation).
2. The current model jointly estimates detection rates and true number of nests, \( P(N_{obs}^{i,s} \mid \hat{P}_i, \hat{N}_s) \), rather than conditionally \( P(N_{obs}^{i,s} \mid \hat{N}_{i,s}) \) for each nest as in the data-augmentation model. The reason is that we have no way to estimate, for the aggregated data, the probability of a nest being present and thus estimate at plot- rather than nest-level.

We can plot the estimated (x-axis) and observed number of nests per plot (y-axis):
Clearly the correction applied is larger in plots with more ant nests, or, phrased differently, the adjustment is proportional. We can plot the posteriors and their mean:

```r
par(mar = c(3, 4, 1, 1), mfrow = c(4, 4))
for (i in 1:16) {
  dens <- antdetectBinom$BUGSoutput$sims.list$Ntrue[, i]
  tdens <- table(dens)
  plot(as.numeric(names(tdens)), tdens/length(dens), las = 1, col = "grey70",
       type = "h", lwd = 3, ylab = "", xlab = "", xlim = c(min(max(0,
                  plots[i, 9] - 1), as.numeric(names(tdens))),
                  max(c(1, as.numeric(names(tdens))))), ylim = c(0,
                  max(tdens/length(dens))),
       axes = F)
  axis(1, at = 0:40)
  points(plots[i, 9], 0, pch = 17, cex = 1.5, col = "red", xpd = T)
  legend("topright", bty = "n", legend = paste("plot", i), cex = 1.2)
  lines(rep(antdetectBinom$BUGSoutput$mean$Ntrue[i], 2), c(0, 1),
        col = "black", lwd = 1.5)
}
```
Histograms are Bayesian posteriors for \( \hat{N}_s \), with mean estimate indicated by the vertical black line, and the red triangle indicates \( N_{\text{obs}} \) for each plot.

6 Plot-level N-mixture estimation (not mentioned in the main document)

Royle (2004) proposed to view observed abundances as mixtures of Poisson distributions, thus extending the patch-occupancy idea to abundances. As is the case for the previous Bayesian plot-level analysis, this approach cannot accommodate nest traits and is thus at the plot level.
library(unmarked)
obs <- matrix(as.character(1:8), 16, 8, byrow = T)
sitecovs <- data.frame(X = as.factor(1:16))
antsumf <- unmarkedFramePCount(y = as.matrix(plots[, -9]), obsCovs = list(observer = obs),
    siteCovs = sitecovs)

(fit <- pcount(~observer ~ X, data = antsumf, K = 50, mixture = "P"))

Call:
pcount(formula = ~observer ~ X, data = antsumf, K = 50, mixture = "P")

Abundance:

| Term     | Estimate | SE     | z      | P(>|z|) |
|----------|----------|--------|--------|---------|
| (Intercept) | 3.760    | 0.269  | 13.9927| 1.73e-44|
| X2       | -0.530   | 0.289  | -1.8360| 6.64e-02|
| X3       | -0.889   | 0.327  | -2.7162| 6.60e-03|
| X4       | -15.401  | 351.669| 0.0438 | 9.65e-01|
| X5       | -0.156   | 0.327  | -0.4603| 6.60e-03|
| X6       | -2.497   | 0.607  | -4.1172| 3.83e-05|
| X7       | -3.589   | 1.015  | -3.5378| 4.03e-04|
| X8       | -1.500   | 0.415  | -3.6128| 3.03e-04|
| X9       | -1.027   | 0.344  | -2.9879| 2.81e-03|
| X10      | -0.352   | 0.275  | -1.2818| 2.00e-01|
| X11      | -3.674   | 1.025  | -3.5846| 3.38e-04|
| X12      | -0.402   | 0.279  | -1.4398| 1.50e-01|
| X13      | -1.931   | 0.494  | -3.9078| 9.31e-05|
| X14      | -1.719   | 0.456  | -3.7726| 1.62e-04|
| X15      | -0.569   | 0.293  | -1.9400| 5.24e-02|
| X16      | -0.438   | 0.282  | -1.5501| 1.21e-01|

Detection:

| Term     | Estimate | SE     | z      | P(>|z|) |
|----------|----------|--------|--------|---------|
| (Intercept) | -1.3132  | 0.326  | -4.034 | 5.49e-05|
| observer2 | 0.0425   | 0.206  | 0.206  | 8.37e-01|
| observer3 | 0.2033   | 0.202  | 1.004  | 3.15e-01|
| observer4 | 0.1246   | 0.204  | 0.611  | 5.41e-01|
| observer5 | 0.3519   | 0.200  | 1.757  | 7.90e-02|
| observer6 | 0.1839   | 0.203  | 0.907  | 3.64e-01|
| observer7 | 0.6556   | 0.202  | 3.248  | 1.16e-03|
| observer8 | 0.7189   | 0.203  | 3.534  | 4.10e-04|

AIC: 449.1964

ests <- fit@estimates@estimates$det@estimates
# observer probabilities:
plogis(c(observer1 = unname(ests[1]), ests[2:8] + ests[1]))

observer1 observer2 observer3 observer4 observer5 observer6 observer7 observer8
0.2119490 0.2191330 0.2478838 0.2335077 0.2766223 0.2442902
0.3412760 0.3556458
Observer rates are estimated lower, and number of nests accordingly higher, than in the Bayesian model. Otherwise the model confirms the results of the previous models.

7 Plot-level maximum likelihood estimation

We can dispense with the Bayesian nature of the previous step as we are using uninformative priors only, and thus in this case use maximum likelihood to estimate the model parameters (observation rates and true number of nests per plot). Again we assume that the number of nests detected by observer $i$ at plot $s$ is a draw from a binomial distribution, i.e. $N_{i,s} \sim \text{Binom}(size = \hat{N}_s, p = \hat{P}_i)$. Thus, across the 16 plots and 8 observers we have to optimise 24 values to find the maximum likelihood fit for the 128 data points in plots.

```r
Nobs <- plots[, -9]  # remove Nmin
# define the maximum likelihood based on the 24 parameters:
toopt <- function(parms, Nobs = plots[, -9]) {
  Ntrue <- exp(parms[1:16])  # ensure positive values
  Pi <- plogis(parms[17:24])  # ensure values in [0,1]
  Nsest <- tcrossprod(Ntrue, Pi)
  -sum(dpois(as.matrix(Nobs), lambda = as.matrix(Nsest), log = T))
}

# use Nmin as start values for Ntrue and 0.5 for detection rates:
Nmin <- plots[, 9]
(op <- optim(par = c(log(Nmin + 1), rep(0.5, 8)), fn = toopt, method = "BFGS", hessian = F))
```

```r
$par
[6] 0.3491841 -0.5671277 1.7842722 2.2362588 2.9141346
[16] 2.8340923 -0.6153202 -0.5634996 -0.3626869 -0.4619341
[21] -0.1688646 -0.3873130 0.2611425 0.3585991

$value
[1] 198.443

$counts
function gradient
   91    43

$convergence
[1] 0
```
Maximum-likelihood estimates are somewhat higher than $N_{\text{obs}}$ (= Nmin), but in plots 2, 6, 9 and 11 they are lower (than the observed number of nests). But the results are extremely similar to the Bayesian plot-level analysis and also compare well to the data-augmentation results (which are essentially identical to $N_{\text{obs}}$):

```r
cor(cbind(mlEst = bestguessNtrue, BayesEst = antdetectBinom$BUGSoutput$mean$Ntrue, dataAug = estimated[, 2]))
```

<table>
<thead>
<tr>
<th></th>
<th>mlEst</th>
<th>BayesEst</th>
<th>dataAug</th>
</tr>
</thead>
<tbody>
<tr>
<td>mlEst</td>
<td>1.0000000</td>
<td>0.9963630</td>
<td>0.9517471</td>
</tr>
<tr>
<td>BayesEst</td>
<td>0.9963630</td>
<td>1.0000000</td>
<td>0.9717419</td>
</tr>
<tr>
<td>dataAug</td>
<td>0.9517471</td>
<td>0.9717419</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>

```r
par(mar = c(5, 5, 1, 1))
plot(estimated[, 2], antdetectBinom$BUGSoutput$mean$Ntrue, xlab = expression(hat(N)[s]^\{augmented\}), ylab = expression(hat(N)[s]), las = 1)
points(estimated[, 2], bestguessNtrue, pch = 16)
abline(0, 1)
legend("bottomright", pch = c(1, 16), legend = c("Bayes estimates", "ML estimates"), bty = "n", cex = 1.2)
```
The figure shows estimates from the site-level Bayesian and maximum likelihood approach (y-axis) against the nest-level data-augmentation results (x-axis). Line gives perfect accordance (1:1).

We think that we can thus use the site-level maximum likelihood approach in lieu of the Bayesian plot-level model, particularly when in the next step we re-run the analysis many times for different combinations of observers. This takes only seconds using the maximum likelihood approach, but would take many hours with the Bayesian plot-level model.

Computing confidence intervals or standard errors for the maximum likelihood estimates is a bit involved, as asymptotic errors (based on the Hessian matrix) are very unreliable for such small data sets. We thus use bootstrapping instead.

```r
# draw, for each plot, with replacement from the nests data set:
plotNames <- substring(rownames(nests), 3, 4)
```
n.bs <- 1000  # number of bootstraps
detProb.bs <- matrix(NA, nrow = n.bs, ncol = 24)
for (n in 1:n.bs) {
plots.bs <- matrix(0, nrow = 16, ncol = 8)
rownames(plots.bs)[c(1:3, 5:16)] <- unique(plotNames)
rownames(plots.bs)[4] <- "04"
for (i in unique(plotNames)) {
  thisPlot <- nests[which(plotNames == i), ]
  plotBS <- thisPlot[sample(nrow(thisPlot), nrow(thisPlot), replace = T), ]
  plots.bs[rownames(plots.bs) == i, ] <- colSums(plotBS)
}
(op.bs <- optim(par = c(log(Nmin + 1), rep(0.5, 8)), fn = toopt,
               Nobs = plots.bs, method = "BFGS"))
detProb.bs[n, 17:24] <- plogis(op.bs$par[17:24])  # detection rates
detProb.bs[n, 1:16] <- exp(op.bs$par[1:16])  # nest estimates
}

# observation rate estimates:
round(colMeans(detProb.bs[, 17:24]), 2)

[1] 0.34 0.36 0.40 0.38 0.45 0.39 0.55 0.58

round(apply(detProb.bs[, 17:24], 2, sd), 3)

[1] 0.059 0.059 0.066 0.064 0.073 0.064 0.095 0.110

round(apply(detProb.bs[, 17:24], 2, quantile, c(0.025, 0.975)), 3)

2.5% 0.246 0.263 0.301 0.279 0.342 0.299 0.418 0.434
97.5% 0.481 0.496 0.564 0.532 0.635 0.635 0.794 0.866

# nest number estimates:
round(colMeans(detProb.bs[, 1:16]), 2)

[1] 27.58 15.92 11.28 0.00 23.27 1.50 0.59 6.15 9.25 19.95
[11] 0.30 18.49 3.79 4.99 15.63 17.77

round(apply(detProb.bs[, 1:16], 2, sd), 3)

[1] 6.177 3.322 2.692 0.000 4.343 0.300 0.086 1.897 2.700 3.853

round(apply(detProb.bs[, 1:16], 2, quantile, c(0.025, 0.975)), 3)

2.5% 19.428 9.375 6.316 0 14.924 0.948 0.409 2.804 4.769
97.5% 42.517 22.337 16.642 0 31.418 2.100 0.744 10.256 15.143
2.5% 12.183 0.205 11.779 1.720 2.148 8.393 11.046
97.5% 27.105 0.372 24.678 7.025 7.921 22.960 24.311
8 Simulating more (and fewer) observers

8.1 How many nests would we have estimated with fewer observers?

It is easy to simulate fewer observers by simply randomly drawing the desired number of observers from the data and repeating the (maximum-likelihood) analysis (the Bayesian would take quite a long time, since we have to repeat this random drawing many times).

```r
k <- 3
Y <- Nobs[, c(1, 3, 5)]
tooptk <- function(parms, k) {
  Ntrue <- exp(parms[1:16])  # ensure positive values
  Pi <- plogis(parms[17:(17 + k - 1)])  # ensure values in [0,1]
  estY <- tcrossprod(Ntrue, Pi)
  -sum(dpois(as.matrix(Y), lambda = as.matrix(estY), log = T))
}
(opk <- optim(par = c(log(Nmin + 1), rep(0, k)), fn = tooptk, k = k,
  control = list(maxit = 2000)))

$par
 [6] -2.4845207 0.6931455 2.7263550 2.8637990 3.8470075
 [16] 3.7855818 -1.7505154 -1.5653785 -1.4278722

$value
 [1] 67.59848

$counts
 function gradient
 10566 NA

$convergence
 [1] 0

$message
 NULL

round(exp(opk$par[1:16]), 1)  # estimated number of nests

[1] 66.8 24.9 18.9 0.2 58.0 0.1 2.0 15.3 17.5 46.9 0.1 48.8
[13] 3.7 9.9 40.5 44.1

round(plogis(opk$par[17:(17 + k - 1)]), 2)  # estimated detProb per observer

[1] 0.15 0.17 0.19

# this can be looped through 1000 times to get error bars for the
# estimates, always drawing a random set of observers; and then we
# repeat this with k=2 to k=8; here the example for k=3:
Nreps <- 10  # only for illustration the value is set low; set this to something larger (takes .3s per rep)
```

29
k <- 3
trueNmat3 <- matrix(NA, ncol = 16, nrow = Nreps)
for (i in 1:Nreps) {
    # choose a random set of observers:
    useTheseObs <- sample(8, k)
    Y <- Nobs[, useTheseObs]
    # compute the trueY based on this:
    opk <- optim(par = c(log(Nmin + 1), rep(0, k)), fn = tooptk, k = 3,
                 control = list(maxit = 50000))
    if (opk$convergence != 0)
        stop("not converged!")
    trueNmat3[i, ] <- exp(opk$par[1:16])
    cat(i, " ")
}

boxplot(trueNmat3, las = 1, ylab = "estimated number of ant nests",
        xlab = "plot", col = "grey70")
Note that these values are much higher when only few observers are used for estimation. Let us briefly try and understand why.

The estimated values of the observed $N_{i,s}$ are the cross-product of the estimated number of nests in a plot, $\hat{N}_s$ and each observers detection probability, $P_i$. If we estimate high values for $\hat{N}_s$, we can ‘compensate’ this by low values of $P_i$. If variability is high (because we have only few observers), we have little to go on for estimating either value. There is now the choice between ‘many nest, poor detection rates’ and ‘few nests, high detection’. As it works out, in the binomial probability mass function, it is easier to accommodate highly variable data with ‘many nests, poor detection’. Thus, the poorer the data (read: the fewer observers), the more the estimation will overestimate the true number of nests.

Repeating this for different values of $k$ will lead to Fig. 3 in the main text. We do not provide the code here, but it is near-trivial to adapt the above for any value of $k$. 

![Plot showing estimated number of ant nests across different plots](image-url)
8.2 How does overall detection rate change with the number of observers like ours?

As detailed in the main paper, we require two different probabilities to simulate more observers: (a) the detection rate $P_i$ of each observer $i$ (which we have from either the maximum-likelihood estimation or the Bayesian analysis); and (b) the probability that a second observer discovers a new (complementary) nest, $P_c$. We compute $P_c$ for each pair of observers, yielding a matrix from which to sample when simulating more observers (or indeed fewer).

```r
# add overlooked species according to jack1 estimate (intermediate # between Chao and jack2)
nestsAll <- rbind(as.matrix(nests), matrix(0, nrow = 25, ncol = 8))
# nests which second observers found but first overlooked
# (quantifying complementarity):
Pc.mat <- matrix(0, 8, 8)
colnames(Pc.mat) <- rownames(Pc.mat) <- colnames(nests)
for (i in 1:8) {
  for (j in 1:8) {
    tt <- table(nestsAll[, i], nestsAll[, j])
    Pc.mat[i, j] <- tt[1, 2]/sum(tt[1, ]) # proportion of 0s turned into 1s
  }
}
round(Pc.mat, 3)  # note that this matrix is (obviously) not symmetric!

# mean Pc value:
mean(c(Pc.mat[lower.tri(Pc.mat)], Pc.mat[upper.tri(Pc.mat)]))  # 0.2565

[1] 0.2568836

hist(c(Pc.mat[lower.tri(Pc.mat)], Pc.mat[upper.tri(Pc.mat)]), main = "",
     xlab = "Pc")
```

01 02 03 04 05 06 07 08
01 0.000 0.159 0.265 0.257 0.283 0.239 0.442 0.496
02 0.144 0.000 0.207 0.225 0.243 0.216 0.450 0.441
03 0.194 0.146 0.000 0.204 0.223 0.223 0.350 0.447
04 0.215 0.196 0.234 0.000 0.290 0.262 0.393 0.477
05 0.147 0.116 0.158 0.200 0.000 0.137 0.358 0.411
06 0.173 0.163 0.231 0.240 0.212 0.000 0.423 0.452
07 0.182 0.208 0.130 0.156 0.208 0.221 0.000 0.351
08 0.219 0.151 0.219 0.233 0.233 0.219 0.315 0.000

# mean Pc value:
mean(c(Pc.mat[lower.tri(Pc.mat)], Pc.mat[upper.tri(Pc.mat)]))  # 0.2565

[1] 0.2568836
# note that this is the complement to the value computed above in joint.mat!!

We now have a matrix with values representing the probabilities of discovering new nests overlooked by one previous observer. The chance of \( k \) observers to all overlooking a nest is \((1 - P_c)^k(1 - P_i)\). We turn this into a little function and plot it for \( k \) from 1 to 10.

```r
overlooked <- function(k, Pc = 0.2565, Pi = 0.42) {
  # Pd is mean of Bayesian estimates returns the probability of
  # having overlooked nests
  (1 - Pc)^((k - 1) * (1 - Pi))
}

par(mar = c(5, 5, 1, 1))
plot(1:10, overlooked(1:10), type = "o", pch = 16, cex = 2, las = 1,
     ylab = "probability of overlooking a nest", xlab = "number of observers",
     lwd = 2, col = "grey40")
abline(h = 0.1, col = "grey", lty = 2)
```
Clearly, and obviously, the more observers we have, the lower is the chance of overlooking a nest. With 8 observers we cross to below 10% overlooked nests.

This plot ignores the variation around the detection and complementarity rates. So we now open the function to bootstrapping (i.e. sampling with replacement) $P_i$ and $P_c$ and run it on the real data ($P_i$ from maximum likelihood).

```r
clearlyB <- function(k, Pc, Pi) {
    # recursive problem!
    if (k == 1)
        return((1 - sample(Pi, 1))) # for one observer: 1-detection probability
    # for two or more observers:
    (1 - sample(Pc, 1)) * clearlyB(k - 1, Pc, Pi)
}

# run a test for 10 observers, 1000 repetitions:
hist(replicate(1000, clearlyB(10, Pc = c(Pc.mat[lower.tri(Pc.mat)],
               Pc.mat[upper.tri(Pc.mat)]), Pi = plogis(op$par[17:24])), main = "10 observers",
       xlab = "proportion overlooked")
```
So we can now compute the overlooking probability for 1 to 10 (or more) observers.

```r
simuObservers <- sapply(1:10, function(x) replicate(1000, overlookedBS(x,
    Pc = c(Pc.mat[lower.tri(Pc.mat)], Pc.mat[upper.tri(Pc.mat)]),
    Pi = c(0.338, 0.35, 0.394, 0.37, 0.439, 0.388, 0.533, 0.561))))
par(mar = c(5, 5, 1, 1))
boxplot(simuObservers, las = 1, whisklty = 1, col = "grey", ylab = "probability of overlooking any nest",
    xlab = "number of observers")
abline(h = c(0.1, 0.05, 0.01), col = "darkgrey", lty = 2)  # targets for accuracy
```
The black lines are the same values as in the previous dot-plot, but now we also get an estimate of the uncertainty around this value.

9 Non-parametric omission error analysis

For completeness, we also include the more traditional way to estimate overlooked nests, building on coarse approximations of the ratio of rare and common events. Each observer’s records are a sample of the true nests at each plot. We can use non-parametric richness estimators to predict the number of nests across all plots. This is akin to having multiple recordings of a community and estimating the total number of species in the pool. The nests-data have to be transposed before analysis to have “species” (i.e. nest locations) as columns.
As boot is frequently reported as underestimating the true richness, we shall only use Chao’s S, jack1 S and jack2 S as estimators of the true number of nests across all plots. Thus, we have sampled 147/164*100% = 90% (Chao), 85% (jack1) or 82% (jack2) of all nests, which suggests a high sampling coverage:

```
par(mar = c(4, 4, 1, 1))
plot(specaccum(t(nests)), xlab = "number of observers", ylab = "number of nests recorded", las = 1, lwd = 2)
```

To visualise which ant nests were detected by who, we can plot the nests data as a matrix.

```
largest <- read.csv("nestSizes.csv", row.names = 1) # read file in again to get all nest sizes
nestSize <- largest$Height
par(oma = c(0, 0, 0, 0.1)) # create space for label to the right
# overlooked species according to jack2 estimate (see further # below):
nests2 <- rbind(as.matrix(nests), matrix(0, nrow = 44, ncol = 8))
visweb(t(nests2), labsize = 5, prednames = F, preynames = F, clear = F)
# here comes the symbol/col for nest size:
points(((1:147) - 0.5)[order(rowSums(nests), decreasing = TRUE)],
       rep(9, 147), pch = 16, col = "darkgrey", cex = 1.3 * ifelse(nestSize == 10, 0.3,
       ifelse(nestSize == 50, 0.6, ifelse(nestSize == 100, 0.9, 1.2))))
```
Dots above the nest indicate the nest’s size. Empty cells to the right were unrecorded by all observers. The best guess for $\hat{N}_{\text{true}}$ of the different methods is indicated by a triangle alongside the method.

Platform, session and package information:

```
sessionInfo()
```

R version 3.2.3 (2015-12-10)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.10.5 (Yosemite)

locale:
attached base packages:
[1] stats  graphics  grDevices  utils  datasets  methods  base

other attached packages:
[1] unmarked_0.11-0  Rcpp_0.12.5  reshape_0.8.5
[4] lme4_1.1-12  Matrix_1.2-6  truncnorm_1.0-7
[7] R2jags_0.5-7  rjags_4-6  coda_0.18-1
[10] bipartite_2.06  sna_2.3-2  vegan_2.3-5
[13] lattice_0.20-33  permute_0.9-0

loaded via a namespace (and not attached):
[1] formatR_1.4  nloptr_1.0.4  plyr_1.8.4
[4] tools_3.2.3  boot_1.3-18  digest_0.6.9
[7] evaluate_0.9  nlme_3.1-128  mgcv_1.8-12
[10] igraph_1.0.1  yaml_2.1.13  parallel_3.2.3
[13] spam_1.3-0  raster_2.4-18  stringr_1.0.0
[16] cluster_2.0.4  knitr_1.13  fields_8.4-1
[19] maps_3.1.0  grid_3.2.3  rmarkdown_0.9.6
[22] sp_1.1-1  minqa_1.2.4  magrittr_1.5
[25] codetools_0.2-14  htmltools_0.3.5  R2WinBUGS_2.1-21
[28] MASS_7.3-45  splines_3.2.3  abind_1.4-3
[31] stringi_1.1.1

The above code is licensed under CC-BY-SA 4.0.