Improving citizen science data for long-term monitoring of plant species in the Netherlands

Arco J. van Strien1 · Jelle S. van Zweden1 · Laurens B. Sparrius2 · Baudewijn Odé2

Received: 26 October 2021 / Revised: 13 May 2022 / Accepted: 20 June 2022 / Published online: 18 July 2022
© The Author(s) 2022

Abstract
In 2012, a new volunteer-based recording scheme for vascular plants was launched in the Netherlands. Its purpose is to track the changes in the number of occupied 1-km grid cells for as many native plant species as possible between survey rounds of 8 years. We did not prescribe a strict field protocol to minimize variation in observer effort, but instead chose to statistically correct for this variation with occupancy models. These models require replicated visits to a grid cell per season, which was implemented by having two independent observers survey grid cells and record all plant species observed. Now that a first survey round has ended (2012–2019), we evaluate our approach, i.e. we tested whether the scheme has the potential to produce proper trend estimates. The number of occupied grid cells in the first round was estimated per species, using an occupancy model with day of year, visit duration and observer experience as covariates for detection. The detection probability, which was 0.43 on average, strongly depended on visit duration and day of year. It was possible to estimate the number of occupied grid cells quite precisely for several hundreds of species, such that the statistical power is expected to be high enough to detect changes of 10% between survey rounds. For rare species, however, the power to detect changes is expected to be quite low. We conclude that the approach works well, but further improvements are suggested.

Keywords Bayesian method · Citizen science · Detection probability · Ecological monitoring · JAGS · Occupancy modelling · Semi-structured survey · Vascular plants
Introduction

Citizen science projects provide valuable information on biodiversity, for they allow data collection at spatial and temporal scales that would be difficult to achieve by other means (Schmeller et al. 2009; Altwegg and Nichols 2019). Most of these data are collected at observer-selected locations without applying standardized field protocols. Therefore, assessing trends in species distribution or abundance using such ‘opportunistic’ data is challenging, because variation in observer effort and non-representative sampling of geographical areas may lead to biased or even spurious trends (Dennis and Thomas 2000; Kuussaari et al. 2007; Brown and Williams 2019).

The monitoring of plant species in the Netherlands illustrates this challenge. Since the beginning of the twentieth century, records of vascular plants have been collected by volunteers who surveyed a grid cell of about 1.3 × 1 km in a year and recorded the presence of every vascular plant species observed (Sparrius et al. 2019; Tamis et al. 2005). There were hardly any further field protocol requirements and no information on the observation process was collected, which made it difficult to assess observer effort. Trend estimation using these data was hence far from perfect, so around 2010 we started thinking about a new monitoring scheme. Our aim was to assess trends in distribution in as many native plant species as possible, by comparing the occupancy, i.e. the proportion of occupied 1-km grid cells, of a species between survey rounds.

One option for the new scheme design was to standardize the field surveys, because a rigorously designed survey scheme with standardized field protocols would result in better trend estimates (Yoccoz et al. 2001; Brown and Williams 2019). But the idea of prescribing a standardized field protocol did not elicit great enthusiasm among the volunteers and the engagement in a fully designed scheme was expected to be low. There was also a more fundamental reason not to opt for a standardized protocol: it is nearly impossible to standardize the field method if entire grid cells of 1-km need to be surveyed. It would imply the choice for many small subunits within the 1-km grid cells to collect enough data for less common species, such as being applied by Chen et al. (2013) and Pescott et al. (2019). In the Netherlands, a large-scale scheme based on small squares surveyed by professional field workers is already applied (Van Duuren et al. 2007) and, indeed, this scheme delivers too sparse data for many species to estimate national trends.

The better option, we believe, was to stick to the 1-km grid cells, to refrain from strict protocols, and instead adopt less far-reaching enhancements. In 2012, we started a new volunteer-based monitoring scheme, in which as many grid cells as possible are surveyed twice per year by independent observers. These observers are encouraged to survey preselected grid cells in order to achieve a proper geographical distribution of surveyed grid cells and a time limit is prescribed to reduce the variation in observer effort. Once a grid cell is surveyed, it is highlighted as having priority for the remainder of the season, so that other observers may sign up for the same grid cell. This within-season replication allows for the application of occupancy models (MacKenzie et al. 2006), which can disentangle detection and occupancy probabilities using records from independent replicated visits. Without such separation between probabilities, higher observer effort over time may result in deceptive positive species trends, whereas it should only affect detection probability. Occupancy models are therefore currently the most powerful tool to adjust for variation in observer effort (Kéry et al. 2010; Van Strien et al. 2011; 2013; Isaac et al. 2014). Recent applications of this idea are assessments of
trends in the distribution of bryophytes, lichens, butterflies, dragonflies and other invertebrates derived from opportunistic data (Van Strien et al. 2013; Outhwaite et al. 2019). To assess trends in as many plant species as possible, a great number of grid cells need to be surveyed. Because it is not feasible to survey many grid cells every year, we have chosen to compare occupancy between survey rounds of several years long, thereby neglecting the yearly variation in occupancy within a survey round. This is the same idea as in many Atlas studies (Kuussaari et al. 2007). Now that a first survey round has ended (2012–2019), we examine how well our approach has worked. More specific, we address three questions: (1) Are species detected with sufficiently high probability and how do common and rare species differ in this respect? (2) For how many species can we detect a decline in the number of occupied grid cells in a virtual second round? (3) Is it possible to further improve the scheme without requiring too much extra field effort from observers? We applied standard single-species occupancy models to the data gathered in the first round to estimate detection and occupancy probabilities. Thereafter, using the model results as ingredients, we assessed the expected power to detect declines. To identify opportunities to further increase the statistical power of the scheme, we also examined survey-specific conditions and plant traits that may affect the detection of species.

Material and methods

Data

Each year of the survey round (2012–2019) a picklist of ~500 1-km priority grid cells was made available through a web application (https://www.verspreidingsatlas.nl/projecten/floron/nem-vaatplanten.aspx). The idea was to provide a set of grid cells evenly spread across the country, yet the selection procedure was not probability-based (Brown and Williams 2019). Observers were encouraged to adhere to this picklist, but they could also add grid cells themselves. As a result, surveyed grid cells were not evenly distributed across the country, but show oversampling of the highly populated western part of the country and undersampling of some other regions (Fig. 1). Each grid cell was only surveyed in a single year. Of the 2851 grid cells that were surveyed in total, 1641 were visited once, 1195 were visited twice, and 15 grid cells were visited three times within that year. These replicated visits were always made by different (groups of) observers, who had no access to the species recorded during the other visit(s). Most visits (96%) were made between May 1 and September 30, with mean date July 17th.

The prescribed field protocol was kept simple. The main rule was that observers needed to survey a grid cell during a single day, with a time limit of minimum 2 and maximum 12 h, depending on the number of habitats present in the grid cell. Mean visit duration was 4.2 h. They were asked to inspect a variety of habitats in the grid cell and to record all species of vascular plants observed, both native and non-native species. Information regarding the observation process was also recorded: date of visit, number of observers, identity of the main observer and exact duration of the visit. Many surveys were performed by experienced observers, experience being deducted from the number of records in the National Database Flora and Fauna (NDFF; www.ndff.nl), which contains almost all vascular plant records made in the Netherlands since the eighteenth century. An alternative measure for observer experience may be the number of unique plant species or families identified by an observer. These alternative measures, however, are strongly correlated with the total
number of observations (log-transformed number of observations vs. number of species, Pearson $r = 0.71$, $P < 0.01$; vs. number of families, $r = 0.72$; $P < 0.01$). Moreover, the number of records better reflects the observers’ experience, effectiveness, and thoroughness in field searching behaviour.

**Statistical model**

To estimate detectability and occupancy of individual plant species, we used a static occupancy model (MacKenzie et al. 2006). Occupancy models require detection/
non-detection data per species and, at in least a number of grid cells, replicated visits are needed within a so-called period of closure. Closure means that the occupancy status of a grid cell (occupied or not) does not change over the survey season (here, year). Replicated visits were always conducted in the same year to meet the closure assumption (Chen et al. 2013). Occupancy models estimate the true occupancy as a latent (i.e. unobserved) state, while accounting explicitly for non-detection (MacKenzie et al. 2006).

Our model estimated occupancy $z_i$, using the observations $y_{ij}$ indicating that the species was detected ($y_{ij} = 1$) or not ($y_{ij} = 0$) during survey visit $j$ ($n = 1$–3 visits) at grid cell $i$. This is described by the following two conditional probability statements:

\[ z_i \sim \text{Bernoulli}(\psi_i) \]  
\[ y_{ij} \sim \text{Bernoulli}(p_{ij} \times z_i) \]

where $\psi_i$ represents the occupancy probability of the species in grid cell $i$, and $p_{ij}$ is the probability of detecting the species in grid cell $i$ during visit $j$. Occupancy and detection probability, in turn, were modelled as

\[ \text{logit}(\psi_i) = \alpha_{\text{occ}} \]

\[ \text{logit}(p_{ij}) = \alpha_{\text{det}} + \beta_1 \times \text{date}_{ij} + \beta_2 \times \text{date}_{ij}^2 + \beta_3 \times \text{visit-duration}_{ij} + \beta_4 \times \text{observer-experience}_{ij} \]

Equation 4 contains several survey-specific covariates to take into account differences between the replicated visits which may affect detection. The $\beta_1$ and $\beta_2$-parameters were the linear and quadratic effects of day of year (i.e. season), $\beta_3$ the effect of visit duration expressed as the log of the number of hours, and $\beta_4$ the effect of the categorical variable observer experience (1: $>50,000$ observation records in the NDFF between the year 2000 and the time of the visit; 0: $\leq50,000$ observations). The covariates date, date$^2$ and visit-duration were standardized before analysis. We also tested the number of observers as a covariate for detection, but effects were not significant and/or negligible. Furthermore, we refrained from making the intercept $\alpha_{\text{det}}$ year-dependent, because we found no effect of calendar year on detection in previous tests and often models did not converge.

The model is formulated in a Bayesian framework using the JAGS language (Plummer, 2009) and the R-package R2jags in R v. 4.1.1 (R Core Team 2018). We used vague priors for all parameters and ran the models with three parallel Markov chains of 30,000 iterations each, discarding the first 24,000 as burn-in and a thinning rate of 18, resulting in a posterior based on 1000 samples for further analysis.

The models were run separately for each species with $>10$ presence records ($n=1244$ species) with all grid cells included. By summing the posterior $z_i$ values over grid cells we estimated for each species the proportion of occupied grid cells, i.e. the occupancy estimate, as well as the total number of occupied grid cells as derived parameters. We used the Gelman-Rubin Rhat statistic to judge if convergence was reached in the occupancy estimate (Rhat < 1.1). In such cases, the other model parameters also converged (Rhat < 1.15) in 93% of all cases; in the remaining cases only one or two of the other parameters did not
converge well. A parameter was considered significant when the 95% Bayesian credible interval (CRI) of the posterior sample of a parameter did not include 0.

**Species traits**

To facilitate the searching behaviour of observers in the future, it is helpful to understand which traits (e.g. ecological or phenological) affect the detection probability (Barata et al. 2018). In addition to including covariates in the model for detection, we examined the associations of detection and several species traits. More specific, we tested the following: (i) Do herbaceous species with a short flowering duration in early spring or late summer (i.e. that stop flowering before the end of May or start flowering in August) have a lower mean detection than other herb species? (ii) Are aquatic species with leaves floating on the water surface easier to detect than species that remain largely below the water surface? (iii) Are species smaller than 25 cm less detectable than taller species? (iv) Do observers favour herbaceous species over tree/shrub species and grass-like species, i.e. the families Poaceae, Cyperaceae and Juncaceae, thus leading to lower detection for the latter? (v) Is detection lower for species which have more look-alikes, i.e. does the species belong to a species-rich genus? Traits i-iv were obtained from BioBase (CBS 2003), although this database is not exhaustive. Duistermaat (2020) was consulted for information on trait v. As detection measure per species, we used the inverse logit of alpha.det which is the detection probability on the mean date, with mean visit duration, for the lowest category of observer experience. Differences in detection between subsets of species were tested using the Kolmogorov–Smirnov two-sample test.

**Power analysis**

Power analysis is usually aimed at optimizing field efforts of a monitoring scheme (Guillera-Arroita and Lahoz-Monfort 2012; Barata et al. 2018; Steenweg et al. 2019), but here we use power analysis as a way to assess how the monitoring scheme is performing. More specifically, we assess for how many species the scheme would be able to detect changes in the number of occupied grid cells after a virtual second round with equal sampling effort. The statistical power is the probability to identify a real change in occupancy, and depends on the sample size, the magnitude of the change, the number of visits per grid cells, the detection probability, the initial occupancy and the chosen significance level of the statistical test used (Guillera-Arroita and Lahoz-Monfort 2012).

The main ingredients of our power analysis are the number of occupied grid cells and its associated uncertainty. These ingredients can be found in the posterior distribution of the number of occupied grid cells, which we obtained from the occupancy analysis per species. Detection probability and number of visits per grid cell are implicitly included in these ingredients. For each species, we applied the following procedure:

- We used the occupancy estimate from the 1000 saved posterior samples for further analysis.
- These 1000 values were lowered by 10% to generate the number of occupied grid cells in a virtual second survey round, as if a 10% decline had occurred. We thus assumed that the same number of 2851 grid cells was surveyed and that the uncertainty in the number of occupied grid cells was similar to the first survey round.
- For each of the 1000 values, the difference in the number of occupied grid cells between survey rounds was tested, using a binomial test and a significance level of 0.05.
- Power was defined as the number of times the test was significant. Thus, if a significant change was found in 800 out of 1000 tests, power is 80%; if found in 600 out of 1000 tests, power is 60%; etc.
- The procedure was repeated for a decline of 20%, 30%, up to 100%. We report the power to detect all of these declines, but for illustrative purposes a decline of 10% receives most attention.

Results

Model convergence

Model convergence was not reached for 214 out of 1244 species and these results were discarded. This lack of convergence is largely due to a low number of observations: two-third of the discarded species had < 50 observation records. In addition, we discarded the results of 63 species with low detection probability estimates ($p < 0.1$), because such estimates may be unreliable (MacKenzie et al. 2006). On the other hand, we did use the results of 14 species with low average detection but with much higher detection earlier in the season, such as in case of *Ficaria verna* (Fig. 2). The 967 remaining species contained 847 native species (which is approximately 59% of the 1432 native species in total in the Netherlands; Sparrius et al. 2014) and 120 non-native species, *e.g.* *Hydrocotyle ranunculoides*.

![Detection probability (± CRI) of Ficaria verna and Erigeron sumatrensis in relation to day of year of a survey visit](image.png)
Detection probability

Some species had a very high probability of detection, $p$, such as Plantago lanceolata ($p=0.97$), and few species (Isatis tinctoria and Liparis loeselii) even had a detection probability very close to 1 (Table S1). The average detection probability of the 967 species was $0.43 \pm 0.01$. Occupancy and detection were not particularly strongly related (Spearman rank correlation $r_s = 0.32$; $P < 0.05$). Some highly detectable species ($p > 0.9$) are present in almost all grid cells, such as Trifolium repens and Urtica dioica (Table S1). An example of a common species with mean $p < 0.3$ is Ficaria verna. Rare species too can have mean $p < 0.3$, like Silene conica, or $p > 0.9$, such as the locally abundant species Anagallis tenella.

In many species, $p$ varies during the season; there was a significant effect of date and/or date$^2$ in 499 species (52%). The seasonal effects were especially strong in herbaceous species that start flowering in March or April and stop flowering in early summer and in species that start flowering in August (Fig. 2). Several of these species are completely undetectable after flowering time, because their above-ground parts disappear entirely (e.g. Myosotis ramosissima).

A longer visit duration resulted in higher detection probabilities for 742 out of the 967 species. On average, detection was considerably higher after a 12 h visit as compared to the mean visit duration of 4.2 h or a 2 h visit (Fig. 3). The average detection probability was higher for common than for rare species, regardless of visit duration (Fig. 3; GLM, effect of rarity $= 0.06 \pm 0.01$, $t = 6.50$, $P < 0.001$). The effect of visit duration was also stronger for rare species ($\leq 500$ grid cells) than for more common species (mean $\pm$ se of beta$^3$-coefficient of rare and common species is $0.71 \pm 0.06$ and $0.50 \pm 0.01$, respectively; Welch two sample t-test, $t = 3.783$, df = 475.1, $P < 0.001$).

Mean detection of all species was slightly higher in the group of more experienced observers as compared to less experienced observers (mean $\pm$ se is $0.49 \pm 0.02$ and

![Fig. 3](image-url) Detection probability in relation to visit duration in rare species (occupancy estimate $\leq 500$ grid cells) and more common species ($> 500$ grid cells). Detection probabilities are species averages.
The more experienced observers obtained a significant higher $p$ for 399 species (41%), in which grassy species, trees and shrubs were overrepresented. Curiously, the group observers with less experience detected 44 species (5%) more often than experienced observers while some of these species are highly detectable, e.g. *Hippophae rhamnoides*.

The trait analysis confirmed that $p$ was lower for herbaceous species flowering in spring or late summer than for species flowering in a different period (Table 1). Furthermore, $p$ was lower for submerged plants than for plants with leaves floating on the water surface, such as *Nymphaea alba* (Table 1). On the other hand, species smaller than 25 cm had no significantly lower detection probability than taller species (Table 1). Furthermore, although less experienced observers showed lower detection of grassy species, trees and shrubs, the trait analysis showed no lower detection for these species groups on average (Table 1), presumably because of the huge variation in detection within each group. As expected, species belonging to a genus with more than 35 species had a significant lower $p$ than species from genera with fewer species (Table 1). The species-rich genera in our data-set are *Carex, Juncus, Rubus, Rosa, Salix, Taraxacum* and *Veronica* (n = 80 species). These are genera with well-known issues in species identification which is reflected by the different scores of the two groups of observers: 52% of the species in these species-rich genera is significantly more often recorded by the experienced observers and only 1% of the species more often by observers with less experience.

### Occupancy

Because many species were missed during a single visit, a second visit augmented the naïve estimate of occupancy of species (average occupancy after 1 visit: 0.18 ± 0.01; after 2 visits: 0.20 ± 0.01), but the model estimate of occupancy is considerably higher.

| Species subset                                      | Number of species | Mean detection (± se) |
|-----------------------------------------------------|-------------------|----------------------|
| Herbaceous species flowering in spring or late summer | 70                | 0.31 ± 0.03*         |
| Herbaceous species flowering in a different period   | 632               | 0.45 ± 0.01*         |
| Herbaceous species < 25 cm                          | 118               | 0.47 ± 0.02          |
| Herbaceous species > 25 cm                          | 514               | 0.44 ± 0.01          |
| Herbaceous species                                  | 702               | 0.43 ± 0.01          |
| Grassy species                                      | 146               | 0.44 ± 0.02          |
| Trees and shrubs                                    | 119               | 0.41 ± 0.02          |
| Submerged species                                   | 29                | 0.38 ± 0.04*         |
| Aquatic species on or above water surface           | 26                | 0.53 ± 0.04*         |
| Species of genera with ≤ 35 species                 | 817               | 0.44 ± 0.01*         |
| Species of genera with > 35 species                 | 80                | 0.40 ± 0.02*         |

Species flowering in spring or late summer (n = 70) were excluded in the comparison of all other subsets, except for the subset of all herbaceous species together (n = 702). Asterisks refer to significant differences in detection between species group (Kolmogorov–Smirnov test, $P < 0.05$). Mean detection is the detection on the mean date, for the lowest category of observer experience, with mean visit duration, averaged across all species of the subset.
(0.31 ± 0.02). As expected, the lower the detection probability was, the greater the gap between naïve estimate and model estimate (Fig. 4). The difference is especially noticeable in species that start flowering in spring or late summer. Given that such species are often completely missed during a visit near the mean date (July 17th; Fig. 2), the model estimate of occupancy can be considerably higher than the naïve estimate.

Power analysis

For 246 out of 967 species, it was possible to detect a decline of 10% in the number of occupied grid cells between survey rounds with a statistical power over 80% (Fig. 5A). Most of these are native species; only five are non-native ones. For 568 species, a decline of 30% would be detectable with 80% power or more, and declines of 50% or more would be detectable for at least 778 species (Fig. 5A). If species would disappear completely between two rounds, this could be statistically detected in all but three species.

The power to detect a decline of 10% is higher for species inhabiting more grid cells (Spearman $r_s = 0.98$; Fig. 5B), i.e. it is easier to detect a decline in common species than in rare species. In addition, a higher detection probability contributes to a higher power for common species (Fig. 5B): species occurring in more than 500 grid cells have a statistical power of 56.7 ± 0.69 if $p < 0.3$ (n = 151), but a power of 84.6 ± 0.86 if $p > 0.3$ (n = 380) (Kolmogorov–Smirnov test; P < 0.05). Several submerged species and species from species-rich genera are among the common species with $p < 0.3$ and thus experience the entailing reduced power. For species that are rare (≤ 500 grid cells), a higher detection does not enhance power (Fig. 5B).
Fig. 5  Power analysis of the monitoring scheme. A Number of species with the power to detect a decline between survey rounds, in relation to the magnitude of the decline expressed as a percentage of occupancy in the first survey round. B Power to detect a 10% decline between survey rounds, in relation to the estimated proportion of occupied grid cells in the first round. Each point represents a species. Square grey symbols represent species with detection probabilities ≤ 0.3. Round black symbols represent species with detection probabilities > 0.3
Power is not reduced for herbaceous species flowering in early spring or late summer, despite their relatively low power to detect 10% decline; flowering early spring and late summer: 54.7 ± 3.3%; n = 70; flowering in a different period: 55.6 ± 1.63%; n = 632; Kolmogorov–Smirnov test; P > 0.05). This is because they are well detected during their flowering period.

Discussion

Performance of the monitoring scheme

The results confirm earlier research that detection of plant species is imperfect (Chen et al. 2009, 2013) and that even highly detectable plant species are sometimes missed (Clarke et al. 2012). Furthermore, the analyses confirm that bias in occupancy is considerable if detection is not taken into account (Fig. 4; MacKenzie et al. 2006; Chen et al. 2013). The seemingly low occupancy estimates of almost all species imply a substantial underestimation of species richness in grid cells (MacKenzie et al. 2006). Detection varies considerably between species. Consequently, it is likely that estimation of community metrics, such as functional and phylogenetic diversity, will be biased if imperfect detection is ignored (Xingfeng et al. 2018).

The approach of two independent visits to grid cells has been quickly adopted by the voluntary observers and the resulting data deliver the opportunity to apply occupancy models, allowing us to adjust for imperfect detection as well as for variation in observer effort and observer experience (Van Strien et al. 2013; Isaac et al. 2014; Johnston et al. 2018). Furthermore, the results of our power analysis suggest that the scheme enables us to detect quite minor changes in the number of occupied grid cells for hundreds of plant species. These models fail in case detection is lower than 0.1 (MacKenzie et al. 2006). Fortunately, detection was sufficiently high for most species, although we suspect that a very low detection probability was the reason for not getting model convergence for a number of species. We conclude that our approach—to refrain from highly standardized field protocols and instead focus on the possibilities to statistically correct for variation in observer effort—seems to have worked well. Kelling et al. (2019) advocate a similar approach, which they call a semi-structured survey design, i.e. a citizen science project which can recruit large numbers of participants while collecting sufficient information to account for variation and bias in the data-collection process by analytical techniques.

Opportunities to improve the scheme

Although we feel encouraged to continue our approach, it is not perfect. Occupancy estimates are sufficiently precise for many species to trace changes of 50% between two rounds, but smaller changes of 10%, for instance, are only detectable for a minority of the species (Fig. 5). Therefore, further improvements should be pursued.

Many of the species with limited power are rare species and for quite some rare species data were even too sparse to run a model. The advocated strategy to raise power for rare species is to survey more grid cells (MacKenzie et al. 2006; Barata et al. 2018). There is, however, a trade-off between including more grid cells in the scheme and the length of survey rounds. The first round took 8 years (2012–2019), whereas we would favour a shorter period of, say, 4 years, to detect changes earlier, which means that we need to measure
more grid cells per year. Then there is also the possibility to increase the number of grid cells specifically in areas with many rare species. However, consequently rare species in other areas will be almost completely missed and such a selection procedure may introduce other unwanted biases in the data. The desire to monitor as many species as possible and at the same time achieve high statistical power and timely trends for all is apparently too ambitious.

For a number of common species, statistical power is limited because of low detection. The advocated strategy for such species is to raise the number of replicated visits per grid cell in order to achieve a higher combined $p$ across all visits (MacKenzie et al. 2006). Again, this is not very attractive, because it comes at the expense of the number of surveyed grid cells. It is more attractive to try to raise $p$ per visit without conducting extra visits.

We identified five opportunities to further improve the quality of the data collection without requiring too much extra field work from observers. First, observers should be encouraged to look under the water surface for vegetation more often. Submerged plants frequently had low detection probabilities or a model that did not converge at all. Better data for this group of species is necessary to improve trend assessment and power.

Second, there is still room for improvement of observer skills. The most experienced observers showed higher detection for many species, including species that are often viewed as more difficult, like grasses, tree and shrub species, and species belonging to species-rich genera. This suggests that training observer skills and increasing their experience could raise the detection for many species (Chen et al. 2009; Ahrends et al. 2011; Barata et al. 2018). Also, cell-phone apps for automatic species identification, such as ObsIdentify, may help to further improve observer skills. At the same time, skilled observers need to be reminded that they sometimes forget to note highly detectable species.

Third, observers could be persuaded to extend their survey to a longer duration. Visit duration has a substantial effect on the detection of many species, especially rare ones (Fig. 3). The mean time spent in a grid cell per visit is now rather short and slight extensions will already produce more species detections.

Fourth, it may be helpful to collect extra information on the search activity of the observers. Most observers nowadays already use a smartphone app with GPS to keep track of their search route and the exact time of observations within a grid cell, which enables new opportunities for occupancy modelling (Altwegg and Nichols 2019; Kelling et al. 2019). If such information is combined with existing habitat maps, for instance, the time spent per habitat type may be retrieved and used as a covariate for detection probabilities of species associated with particular habitats. This would be a more subtle covariate than time spent in the entire grid cell, as in our current model.

Fifth, the distribution of surveyed grid cells over the country is clearly imperfect due to the preferential additions of grid cells by observers (Fig. 1), but such geographical bias should be avoided and/or adjusted for as much as possible. To get an idea of how big the geographical bias was, we compared our results to the National Database Flora and Fauna (NDFF) of the Netherlands, which contains > 40 times as many plant records as our scheme. The database contains records from 1975 to 2019 of all grid cells, most of which have been surveyed a number of times over that long period, so that detection approaches 1 in many grid cells. Hence, this lowers detection bias and allows for comparison with our occupancy estimates, which shows that the estimated number of occupied grid cells of the first survey round strongly correlates with the number of grid cells in 1975–2019 (Spearman rank correlation; $r_s = 0.92; P < 0.05; n = 967$). The good match suggests that preferential sampling in our scheme did not induce strongly biased
inferences on occupancy, despite the comparison not being perfect either, since a different time period is covered. Nevertheless, it is better to reduce any bias as much as possible. One way is to fully adopt a probability-based selection procedure and to discourage observers to survey other grid cells than the preselected set. Another possibility is to mitigate bias by filtering out grid cells from oversampled areas or by using model-based methods to adjust for geographical bias (Geldmann et al. 2016; Altwegg and Nichols 2019).

**Model assumptions**

Four model assumptions need to be addressed. First, grid cells of 1 km² are too large and often contain too many different habitat types to be thoroughly surveyed during a visit of several hours. Visits by different observers likely cover different parts of the grid cells to some degree and may also cover different habitats, leading to certain (groups of) species being missed completely in one or the other visit. In our approach, the non-occurrence of the species is interpreted as a non-detection, but this may induce bias in occupancy estimation. However, Kendall and White (2009) demonstrated that the bias that this may induce is negated if the subunits within a grid cell can be considered as surveyed with replacement. And because observers do not have information about the behaviour of other observers of the same grid cell, we believe that our situation is comparable to sampling with replacement, so that the incomplete surveys do not lead to biased estimates.

Second, we assumed that occupancy probability was constant across all grid cells (MacKenzie et al. 2006). This assumption, however, is likely violated because grids cells may be unsuitable for particular species. According to MacKenzie et al. (2006), the violation of this assumption does not necessarily lead to biased estimates of occupancy, but if unsuitable grid cells were filtered out from the analysis, the power of the binomial test to detect a trend is higher. In other words, our power assessments may be considered conservative.

Third, both false negative errors and false positive errors occur in the scheme. We dealt with false negative errors (non-detections) through the application of occupancy models. But we assumed that the data contained no false positive records, although it is obvious that plant species are frequently misidentified, e.g. in the case of species of species-rich genera which may be confused by observers with less experience. It may be useful to include the existence of false positives in future modelling work (Johnston et al. 2022).

Finally, we neglected any variation in detection probability between years. Such unmodelled heterogeneous detection probabilities can lead to occupancy estimates that are biased low (MacKenzie et al. 2006). Lower occupancy estimates are associated with a reduced power (Fig. 5A). Thus, if it were possible to include an annual detection probability in the model specification, the power to detect trends may be higher. Unfortunately, our data were not appropriate to include an annual detection probability in the model. Each grid cell has only been surveyed in one year and as a result for many species there are simply too few detection records to estimate an annual detection probability. After a second scheme round, it will become possible to incorporate a round-dependent detection probability in the model. That enables to correct for systematic changes in observer effort and skills between survey rounds (Kéry et al. 2010; Van Strien et al. 2013).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10531-022-02457-y.
Acknowledgements Leo Soldaat was helpful in the development of the procedure to assess statistical power. We thank Tom van der Meij, Willy van Strien and Marnix de Zeeuw for critically reading the draft. This work would not have been possible without the participation of many voluntary field workers.

Author contributions BO, LBS and AJvS conceived and designed the survey scheme. LBS and BO maintained all communications with the voluntary observers. LBS and BO maintained the database. AJvS and JSvZ analyzed the data and wrote the manuscript.

Funding This project was carried out within the Network Ecological Monitoring (NEM), funded by the Dutch national and provincial governments.

Data availability The relevant data has been submitted to a repository on Dryad, https://doi.org/10.5061/dryad.xksn02vh4.

Declarations

Conflict of interest The authors declare that there were no conflicts of interests for any of the authors regarding this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahrends A, Rahbek C, Bulling MT, Burgess ND, Platts PJ, Lovett JC, Wilkins Kindembha V, Owen N, Ntemi Sallu A, Marshall AR, Mhoro BE, Fanning E, Marchant R (2011) Conservation and the botanist effect. Biol Cons 144:131–140
Altwegg R, Nichols JD (2019) Occupancy models for citizen-science data. Methods Ecol Evol 10:8–21
Arco J, van Strien Chris A, M, van Swaay Marc, Kéry (2011) Ecological Applications 21(7) 2510–2520. https://doi.org/10.1890/10-1786.1
Barata IM, Griffiths RA, Ridout MS (2018) The power of monitoring: optimizing survey designs to detect occupancy changes in a rare amphibian population. Sci Rep 7:16491
Brown ED, Williams BK (2019) The potential for citizen science to produce reliable and useful information in ecology. Conserv Biol 33:561–569
CBS (2003) Biobase 2003. Register biodiversiteit. CBS. Voorburg/Heerlen (available upon request).
Chen G, Kéry M, Zhang J, Keping M (2009) Factors affecting detection probability in plant distribution studies. J Ecol 97:1383–1389
Chen G, Kéry M, Plattner M, Ma K, Gardner B (2013) Imperfect detection is the rule rather than the exception in plant distribution studies. J Ecol 101:183–191
Clarke KD, Lewis M, Brandle R, Ostendorf B (2012) Non-detection errors in a survey of persistent, highly-detectable vegetation species. Environ Monit Assess 184:625–635
Dennis RLH, Thomas CD (2000) Bias in butterfly distribution maps: the influence of hot spots and recorders home range. J Insect Conserv 4:73–77
Duistermaat L (2020) Heukels’ Flora van Nederland. Noordhoff, Groningen, Utrecht
Geldmann J, Heilmann-Clausen J, Holm TE, Levinsky I, Markussen B, Olsen K, Rahbek C, Tøstrup AP (2016) What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers Distrib 22:1139–1149
Guillera-Arroita G, Lahoz-Monfort JJ (2012) Designing studies to detect differences in species occupancy: power analysis under imperfect detection. Methods Ecol Evol 3:860–869
