Diversity and Distributions, (Diversity Distrib.) (2016) 22, 1139–1149

BIODIVERSITY RESEARCH

What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements

Jonas Geldmann[1]*, Jacob Heilmann-Clausen[1], Thomas E. Holm[2], Irina Levinsky[3], Bo Markussen[4], Kent Olsen[5], Carsten Rahbek[1,6] and Anders P. Tøttrup[1]

[1]*Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen E, Denmark,* [2]*Department of Bioscience, Aarhus University, Grenaavej 14, 8410 Roende, Denmark,* [3]*Dansk Ornitologisk Forening - BirdLife Denmark, Vesterbrogade 140, 1620 Copenhagen V, Denmark,* [4]*Laboratory for Applied Statistics, Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen E, Denmark,* [5]*Natural History Museum Aarhus, Wilhelm Meyers Allé 210, 8000 Aarhus C, Denmark,* [6]*Imperial College London, Silwood Park, Buckhurst Road, Ascot, Berkshire, SL5 7PY, UK*

*Correspondence: Jonas Geldmann, Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen E, Denmark.
E-mail: jg794@cam.ac.uk

ABSTRACT

Aim To understand how the integration of contextual spatial data on land cover and human infrastructure can help reduce spatial bias in sampling effort, and improve the utilization of citizen science-based species recording schemes. By comparing four different citizen science projects, we explore how the sampling design's complexity affects the role of these spatial biases.

Location Denmark, Europe.

Methods We used a point process model to estimate the effect of land cover and human infrastructure on the intensity of observations from four different citizen science species recording schemes. We then use these results to predict areas of under- and oversampling as well as relative biodiversity 'hotspots' and 'deserts', accounting for common spatial biases introduced in unstructured sampling designs.

Results We demonstrate that the explanatory power of spatial biases such as infrastructure and human population density increased as the complexity of the sampling schemes decreased. Despite a low absolute sampling effort in agricultural landscapes, these areas still appeared oversampled compared to the observed species richness. Conversely, forests and grassland appeared undersampled despite higher absolute sampling efforts. We also present a novel and effective analytical approach to address spatial biases in unstructured sampling schemes and a new way to address such biases, when more structured sampling is not an option.

Main conclusions We show that citizen science datasets, which rely on untrained amateurs, are more heavily prone to spatial biases from infrastructure and human population density. Objectives and protocols of mass-participating projects should thus be designed with this in mind. Our results suggest that, where contextual data is available, modelling the intensity of individual observation can help understand and quantify how spatial biases affect the observed biological patterns.

Keywords
biodiversity hotspots, citizen science, conservation priority, point process model, species richness, volunteer.

A Journal of Conservation Biogeography

Diversity and Distributions

© 2016 John Wiley & Sons Ltd

DOI: 10.1111/ddi.12477
http://wileyonlinelibrary.com/journal/ddi

INTRODUCTION

The engagement of volunteers in data collection (i.e. citizen science) has the potential to provide unparalleled amounts of data over large temporal and spatial scales, as well as help encourage the public to participate in the scientific process (Tewksbury *et al.*, 2014; Bela *et al.*, 2016). Today, as much as 45% of the ca. 570 million records housed in the Global Biodiversity Information Facility (2015) have been collected by volunteers. Moreover, it has been estimated that the majority of existing natural history data has been collected by amateur volunteers who are not employed at universities or natural history museums (Bell *et al.*, 2008; Mackechnie *et al.*, 2011; Preston, 2013; Roy *et al.*, 2014). Many citizen science projects have a long history (e.g. bird ringing), with data historically stored in personal or society logs. This has made them difficult to access for scientific and management purposes. However, the development of web-based recording with a user-friendly interface and associated databases has resulted in data being increasingly consolidated and available to the research community (Dickinson *et al.*, 2012). Likewise, tools such as Google Earth and dedicated smartphone applications (APPs) using internal clocks and GPS to record time and place have encouraged an even larger group of people to engage in the collection of potentially useful data (Wiggins & Crowston, 2011; August *et al.*, 2015). Where these datasets have a sufficient spatial and temporal resolution, they represent a cost-effective tool for monitoring biodiversity (Schmeller *et al.*, 2009). This has led to volunteer-based data increasingly being used as part of the national reporting towards international targets (Gregory *et al.*, 2005; Tulloch *et al.*, 2013) as well as in research (Powney & Isaac, 2015).

However, the inclusion of citizens in data collection does not come without a cost, and most data collected by volunteers violate one or more fundamental principles of sound experimental design. Isaac *et al.* (2014) identified four categories of sampling biases related to variation in recorder activity: (1) uneven recording intensity over time, (2) uneven spatial coverage, (3) uneven sampling effort per visit and (4) variable abilities to detect species among volunteers. If not addressed, these can lead to unsubstantiated estimates of species richness or changes in abundance, overwriting any real signal in the data, and leading to biased results or in worst cases false conclusions. However, appropriate analysis of volunteer-collected data can help overcome much of this bias (Schmeller *et al.*, 2009; Bird *et al.*, 2014; Powney & Isaac, 2015). Traditionally, sampling biases have been addressed using methods to exclude (filter) and/or standardize subsets of the data, keeping only the reliable observations (e.g. Hickling *et al.*, 2006; Van Calster *et al.*, 2008; Carvalheiro *et al.*, 2013) but at the cost of losing substantial amounts of data (Isaac & Pocock, 2015). Likewise, methodological advances have allowed for modelling species distribution patterns while accounting for uneven detectability (Bird *et al.*, 2014; Isaac *et al.*, 2014). These include using species accumulation curves to address uneven sampling (Heilmann-Clausen & Læssøe, 2012; Eskildsen *et al.*, 2015), inferring sampling effort from the number of species recorded at individual visits (Barnes *et al.*, 2015) or occupancy modelling, where the same sites are visited multiple times (Kery *et al.*, 2010; van Strien *et al.*, 2013). Such analyses offer a powerful toolbox for addressing the various biases inherent in volunteer-collected data. However, these approaches are based on assumptions about the behaviour of the collectors and are extracted from the species recording scheme itself. Thus, these analyses do not incorporate independent contextual data linked to expected spatial biases in the intensity of observations. Spatial patterns in observation intensity originate from two well-known conditions. One, individuals of any species are not randomly distributed across the landscape (Peterson *et al.*, 2011; Erb *et al.*, 2012). Two, the collectors' observations are not randomly distributed in space (Isaac *et al.*, 2014; Powney & Isaac, 2015). Both sources are of concern for the validity of collection schemes based on presence-only data from volunteers (Crall *et al.*, 2011; Bird *et al.*, 2014). When secondary data on sources of bias are available, *a posteriori* expectation of how they impact the behaviour of volunteers can be used to improve current models. This allows for filtering out the effect of such biases rather than filtering the observations. We know that people tend to record close to where they live (Luck *et al.*, 2004; Luck, 2007), in places they enjoy spending time (Hörnsten & Fredman, 2000), and in places known for their biodiversity value (Prendergast *et al.*, 1993). Also, landscape properties, making some areas more accessible and other impenetrable, affect people's behaviour leading to differences in observation intensity. For example, roads have been shown to affect the frequency of plant observations, so that the number of records increased as distance to roads decreased (Kadmon *et al.*, 2004; McCarthy *et al.*, 2012). Similar patterns have been observed for birds (Hanowski & Niemi, 1995; Keller & Scallan, 1999). While such biases are intuitive and acknowledged, existing analyses of volunteer-collected data do not explicitly quantify and incorporate them (Isaac *et al.*, 2014). Thus, including contextual data on well-known biases in models of unstructured volunteer data represents a novel and important contribution to citizen science and can help better utilize existing and new data on species occurrence.

To address this gap in modelling volunteer-collected data, we use a point process model (PPM) to investigate how differences in land cover and spatial bias from human infrastructure affect the distribution of observations across citizen science projects. We examine the intensity of observations (i.e. sampling effort), which is fundamentally different from modelling species richness or abundance. We use these results to investigate how the design and implementation of citizen science schemes affect the observed spatial bias, by examining four citizen science projects, which vary in (1) taxonomic breadth, (2) complexity of the design and (3) number of participants. To our knowledge, this is the first analysis to quantify how variability in sampling design,

ranging from untrained amateurs with no prior training to highly specialized amateur experts, affects the contribution of spatial bias. Finally, we use this information to assess areas of under- and oversampling based on the correlation between corrected sampling effort and recorded species richness. Such maps, while not a direct measure of species richness, can provide a powerful and novel tool for conservation managers who need to make real-time decisions with imperfect data.

METHODS

Study area

Our study was conducted in Denmark, located between Fennoscandia and mainland Europe. Denmark has an estimated species richness of 35–40,000 multicellular species. The total land area is 42,916 km^2 with a total population of 5.6 million people (population density: 131 people km^{-2}). The population is primarily urban with 87.5% living in cities (Danmarks Statistik, 2015). The country is one of the most intensely farmed in the world with more than 62% of the total land area devoted to agriculture (primarily high intensive practices). Forests cover 12% with more than 95% being production forest and < 2% is set aside as non-intervention reserves. Cities and roads make up 10%, while grassland and heaths cover around 9% (Normander *et al.*, 2009). Denmark has a large network of protected areas covering 18% (IUCN and UNEP-WCMC, 2015); however, the vast majority are seminatural or culturally important habitats. Natural habitats are extremely patchy with very few larger connected areas, and across all ecosystems, the state of Danish nature is predominantly poor or unknown (Ejrnæs *et al.*, 2011).

Volunteer biodiversity datasets

To assess the spatial biases in data collected by volunteers, we compiled point-based biodiversity data from four major citizen science projects in Denmark: two taxon-specific schemes recording all species of birds (Bird atlas III) and macrofungi (Svampeatlas), respectively, an all-taxon species recording scheme (Naturbasen) and a simpler scheme covering 30 specific indicators (Biodiversitet Nu). Together, these represent four different participant recruitment strategies, with huge variation in taxonomic breadth and complexity of the sampling design (Table 1).

'Svampeatlas' henceforth referred to as 'Fungal atlas' (www.svampeatlas.dk) ran from 2009 to 2013 with the aim to collect data (distribution and ecology) of all fruit body forming Basidiomycota in Denmark. The project was a collaboration between the Natural History Museum of Denmark, the Danish Mycological Society and MycoKey (http://www.mycokey.com/). The project attracted a variety of volunteers, from trained biologists with specific training in mycology to eager amateurs primarily interested in edible fungi. All records were validated by paid experts, either through photographic documentation or when needed by examination of physical specimens. Two field camps were organized annually based on data gaps to improve coverage, and volunteers were encouraged by competitions to collect in areas with low coverage. Data were recorded via a data-entry portal on the project website.

Bird atlas III henceforth referred to as 'Bird atlas' (www.dofbasen.dk/atlas), the third Danish breeding Bird atlas, runs 2014–2017 and aims to collect data on the distribution of all breeding birds in Denmark at a 5 × 5 km grid. The project is managed by Dansk Ornitologisk Forening (DOF) – BirdLife Denmark, and the fieldwork is carried out by volunteer amateur ornithologists. Observations are recorded online, per grid cell or point locations. Data used in this article cover only the latter, corresponding to 43% of the data collected in the first two breeding seasons, 2014 and 2015. In this period, three atlas camps were held to cover gaps in the coverage. Data are recorded via a data-entry portal on the project website (compatible with mobile devices) and are validated on several scales by experienced volunteers.

The database 'Naturbasen' henceforth referred to as 'All species' (www.fugleognatur.dk) was established in 2001 as the first online, countrywide database for the collection of biodiversity data in Denmark. Citizens can report more than 39,000 species using a dedicated APP (since 2012), which automatically records location and time of an observation, or alternatively through a web-based module at a home computer (since 2001). Records are continually validated by the aid of photographic documentation and a panel of mainly amateur experts. Participants are not encouraged to collect any specific data.

Table 1 Overview of individual datasets.

Dataset	Years	# of obs	# of species	# of recorders	Mean records	Median records
Fungal atlas	2009–2013	292,022	3934	445	657	8
Bird atlas	2014–2015	92,200	207	1061	87	20
All species	2009–2014	429,300	11,581	3682	117	2
Common indicators	2015	46,018	30	6090	8	3

Years refer to the number of years for which data were extracted. This is not always the same as the duration of the citizen science project. For the Common Indicator, scheme number of species refers to the number of indicators which in some cases cover multiple species. Mean and median records refer average number of records per recorder estimated as the mean or median, respectively.

'Biodiversitet Nu' henceforth referred to as '*Common indicators*' (www.biodiversitet.nu) was designed for volunteers with no prior species identification skills. It is based on 30 selected indicators, which are professed to be unmistakable [e.g. a hare (*Lepus europaeus*), any true dragonfly (Anisoptera sp.)]. Guidance is given to the recorder in the form of short text and pictures of the target species and potential confusion species. There is no verification of observations. Data are collected using a dedicated APP which records both location and time or alternatively through a web-based module where the recorder subsequently can record their observations on their home computer.

Spatial data describing spatial bias in effort

Three features were used to account for spatial bias in sampling effort: (1) roads, (2) human population density and (3) land cover. Data on roads were extracted from the National Survey and Cadastre of Denmark (Geodatastyrelsen og Danske kommuner, 2014) as line segments using ARCGIS 10.2. Separate layers were generated for (1) highways, (2) roads broader than 6 m across, (3) roads smaller than 6 m across and (4) footpaths, as these types of roads are expected to influence recording activity differently. A total of 1,630,528 line segments of roads were extracted. For all four types, the shortest Euclid distance between any observation point and the nearest road segment was calculated, generating four distance maps. Human population density at a resolution of 100 × 100 m was extracted from the national database. As this layer was created based on register data from all Danish municipalities, it is extremely accurate. Land cover types were classified using a map consisting of 36 primary land cover classes at a 10 × 10 m resolution (Jepsen & Levin, 2013), which we reclassified in to 18 classes to reduce redundancy and increase clarity (see Fig. S1 and Table S1 in Supporting Information). This map was based on source data from five different geo-datasets for Denmark, ranging from land use types, through maps of agricultural land use to national topographic data (For detailed information on data sources, types and methods, see Jepsen & Levin, 2013). The 18 categories cover 10 different categories of natural habitats, four urban, three classes of agriculture as well as undefined pixels. All datasets were projected using UTM zone 32N, which is the zone covering the majority of Denmark.

Analytical framework

The statistical analyses by PPMs were conducted in R 3.2.2 (R Development Core Team, 2015) using the 'Spatstat' package (Baddeley & Turner, 2005). PPMs describe the number of points in a given area as an outcome of a Poisson distributed random process with the intensity given as the area integral of the underlying intensity field (Baddeley *et al.*, 2015). This intensity field may be related to spatial covariates, and conditional on these covariates, the points are stochastically independent. Thus, this class of models is useful where the object of interest is the location of a point (e.g. measures of abundance or density of records per unit area) in relation to spatial covariates (Baddeley *et al.*, 2015; Renner *et al.*, 2015). Further, it has been argued that PPMs are less sensitive to the effect of scale than analysis based on a predefined grid size (Warton & Shepherd, 2010; Renner *et al.*, 2015). However, while PPMs can be extremely powerful, they are not without their limitations. Kéry & Royle (2016) caution the use of PPMs in particular in cases where measurement errors (false positives and false negative) are large or where points represent a moving object. While both of these issues at first glance seem relevant to our data, it is important to distinguish between the 'observation event' and the 'information' such an observation event represents. In this study, we model the intensity of observations (e.g. the aim is to investigate what spatial factors determine where people record). The information recorded in the observation can be wrong (e.g. a misidentified butterfly or a missed hare), but that does not affect whether or not an observation event did occur, nor that the intensity of observation events has a spatial pattern related to contextual factors. Likewise, while an observation may seek to record a mobile object, the observation itself is not mobile. PPMs therefore represent a powerful tool for understanding the spatial patterns that determines the likelihood of an observation as a function of a series of covariates. The Poisson intensity λ of a point pattern (μ) is modelled as

$$\lambda(\mu) = \beta_1 \mathrm{LC}(\mu) + \beta_2 \log(\mathrm{HPD}(\mu)) + \beta_3 R_{hw}(\mu) + \beta_4 R_{\mathrm{large}}(\mu) + \beta_5 R_{\mathrm{small}}(\mu) + \beta_6 R_{\mathrm{path}}(\mu),$$

where βs represent the effect of the six covariates: (1) land cover LC, (2) human population density HPD, (3) distance to highways R_{hw}, (4) distance to roads > 6 m R_{large}, (5) distance to roads 3–6 m R_{small} and (6) distance to foot paths R_{path}. A best fit model was selected based on Akaike's information criterion (AIC) using a stepwise reduction from the full model (Burnham & Anderson, 2002). Models were inspected using a Pearson residual field and influence plot. To estimate the variance partitioning between the full model and models only containing either land cover or roads and human population density, a McFadden's pseudo-R^2 was calculated for each of the four datasets (McFadden, 1974).

For the *Fungal atlas*, *Bird atlas* and *All species* datasets, we ran five submodels for each set using 46,018 randomly selected observations to assess potential effects of the variable size of the datasets across the four schemes. Results from submodels were consistent with the full models (see Appendix S1 in supporting information online material).

Point process model assumes independence among the points conditional on the covariates. This assumption may be violated in two ways. One, observations from the same human observer may be correlated, for example some observers may report more observations than other. Two, observations may be spatially correlated, for example some areas

may be richer in species and/or more frequently visited. Although the first violation might be modelled using a random effect of observer, this will be ignored due to computational limitations. The second violation was investigated by spatially aggregated residuals. Species richness was calculated in 5 × 5 km grid cells for each of the three datasets which recorded all species within their respective taxa (i.e. *Fungal atlas*, *Bird atlas* and *All species*). To assess how spatial biases in sampling efforts affect species richness measures, we investigated the correlations between the residuals of the full model (see Fig. S2) and the species richness for the four datasets (see Fig. S3) in each 5 × 5 km cell. This gives us a measure of the correlation between sampling efforts and observed species richness. We used these maps to access potential biodiversity hotspots (i.e. areas with high species richness measures despite undersampling) as well as places of potentially low species richness (i.e. areas with low species richness despite extensive sampling).

RESULTS

Agriculture affected the intensity of observations negatively, with more intensive agriculture having a larger negative impact for all four datasets (Table 2). Human-modified land cover classes consistently increased the likelihood of observation, except for areas of resource extraction, where results varied between datasets. For all land cover types classified as natural habitat, the signal was more variable between datasets. As expected, the *Fungal atlas*, collected by a smaller group of specialized volunteers, showed a stronger increase in the intensity of observations in most natural habitats compared to the other datasets, while the *common indicator* dataset and the *Bird atlas* more often experienced decreasing intensities in natural land cover classes, suggesting that they were more heavily affected by sampling bias or that recorded species had lower affinities for specific habitat types (Fig. 1).

Intensity of observation was significantly affected by human population density for all datasets. The *All species* and *common indicator* datasets experienced the largest increases in observation intensity with increasing human population density, while the *Fungal atlas* decreased in observation intensity with increasing human population density (Fig. 2). Effects of roads were more equivocal between datasets (see Fig. S4). For all datasets, highways had a small but significant effect, so that intensity decreased with increasing distance to highways while the intensity increased with increasing distance to roads between 3 to 6 m, the effect being strongest for the *Fungal atlas* followed by the *Bird atlas*, *All species* and the *common indicator* dataset (see Fig. S4).

We calculated McFadden's pseudo-R^2 for the full model, as well as a model containing only the land cover variable (henceforth referred to as nature) and a model containing only roads and population density alone (henceforth referred to as bias). Total R^2 values were highest for the *Fungal atlas* followed by the *Common indicators*, *All species* and the *Bird atlas* (Fig. 3a). When partitioning the explained variance between nature and bias, the contribution of spatial bias increased as the complexity of the volunteer scheme decreased. When relating this to the number of participants contributing observations to the individual schemes, we found that the spatial bias increased significantly with an increasing number of participants (estimate = 0.007, $t = 6.34$, $P = 0.02$; Fig. 3b).

Table 2 Overview of model results for the four datasets.

Dataset	R^2 full	R^2 nature	Mean intensity (km^{-2})	Most likely land cover	Least likely land cover
Fungal atlas	0.297	0.285	6.81	Forest	Intensive agriculture
Bird atlas	0.168	0.155	2.11	Lakes	Intensive agriculture
All species	0.148	0.133	9.71	Dry grassland	Intensive agriculture
Common indicators	0.209	0.183	1.07	Parks and sports	Intensive agriculture

R^2 is the McFadden R^2 value for the best fit model (full) and the model only containing the land cover variable (nature). Mean intensity represents the mean number of observations per km^2. Most likely land cover describes the land cover class with the highest intensity of observations, while least likely land cover represents the land cover class with the lowest intensity.

We investigated areas of high and low sampling effort by comparing the residuals of the intensity of observation (e.g. over- or undersampled areas) with the uncorrected species richness for the *All species*, *Bird atlas* and *Fungal atlas* datasets. Patterns varied between the three datasets. However, for all datasets, areas around major cities showed a combination of very high sampling effort and high species richness (Fig. 4). The *Bird atlas* had higher homogeneity between sampling effort and species richness, while both the *Fungal atlas* and the *All species* datasets had more areas of low sampling effort and low species richness, which suggest areas where more resources are needed to assess the true species richness (Fig. 4).

DISCUSSION

Simplicity amplifies the effect of noise

Our results demonstrate a clear trade-off between the number of participants involved in collecting data and the magnitude of spatial biases: data from citizen science projects with a low number of participants were less affected by roads and human population density than schemes with many participants. This is likely because the *All species* dataset and in particular the *Common indicator* datasets have a higher proportion of ephemeral participants who are more likely to report in areas where they live or commute. Further, these

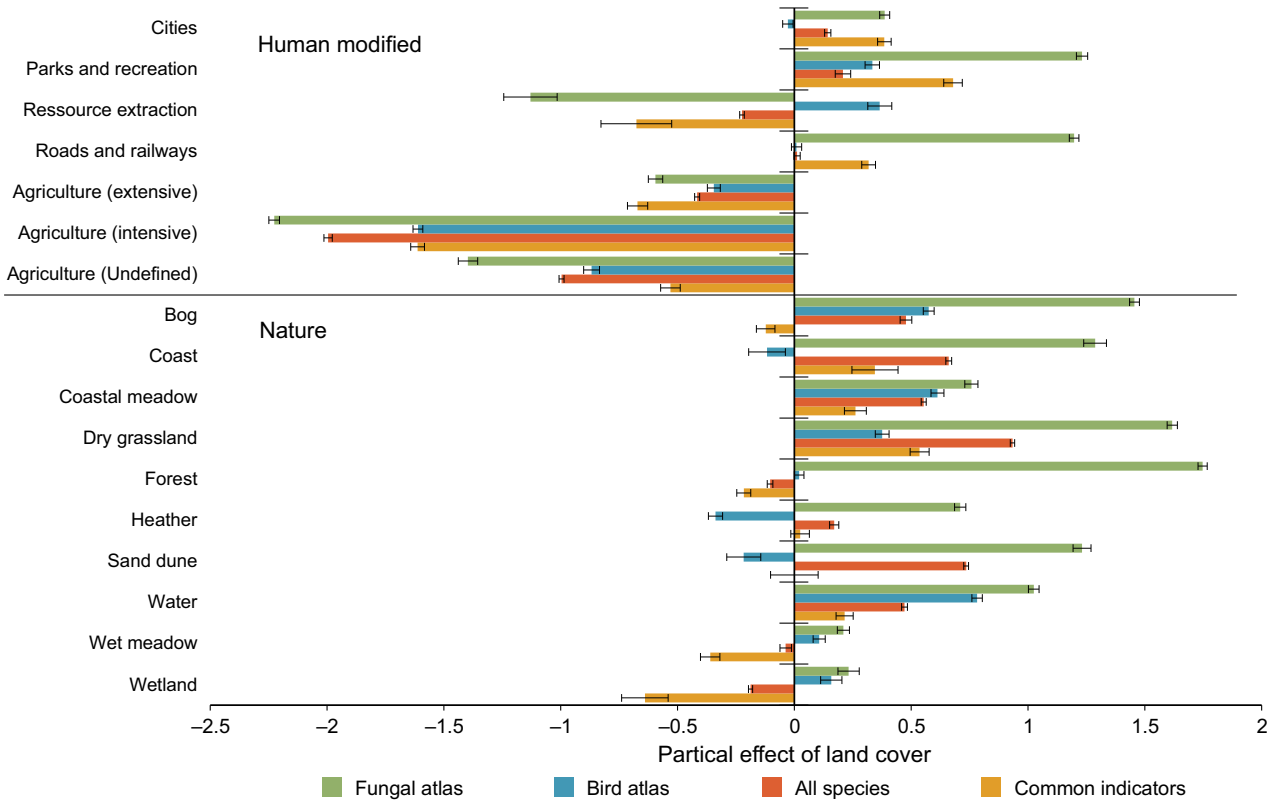

Figure 1 The effect size of the modeled intensity for different land cover classes for the Fungal atlas (green), Bird atlas (blue), All species (orange) and Common indicators (yellow). Positive values indicate increased observation intensity, while negative values indicate decreased observation intensity. Error bars are standard errors.

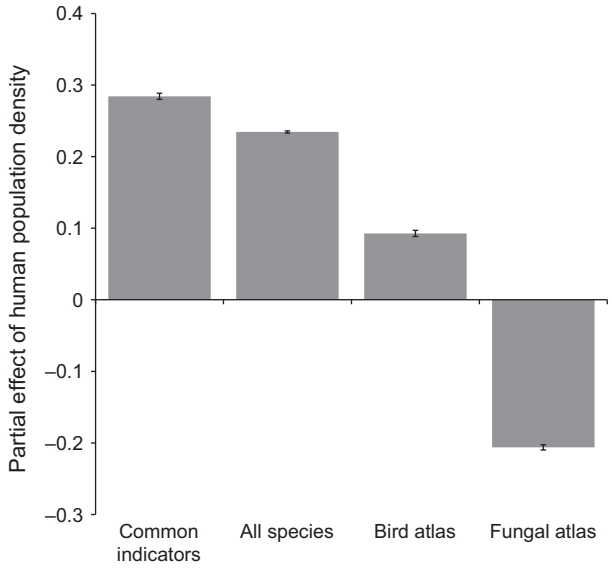

Figure 2 Partial effect of human population density measured as log(number of people per 100 m^2) on the four datasets. Increased human population density affected the density of observations negatively for the Fungal atlas, while the three others increased with human population density. Error bars are standard errors.

two schemes are less explicitly aiming to achieve a thorough coverage or assess species richness of particular taxa but encourage people to report whenever they see something of interest. Conversely, the *Bird atlas* and *Fungal atlas* have been created to explicitly assess the species richness of their respective taxa engaging a smaller group of core participants.

Our results caution against the use of mass-participation schemes without consideration of the trade-offs between increased amounts of data and the value of the individual observation. To what extent these biases undermine the use of volunteer-collected data depends greatly on the objectives of the individual study. For example; studies of phenology or range shifts over time related to climate change may be less affected by over- and underreporting (Dickinson *et al.*, 2012) and thus less sensitive to the participants' skill levels. Likewise, selection of species indicators included in projects can help reduce biases related to misidentification and observability (Gardiner *et al.*, 2012). More importantly, we show that where secondary data on biases exist, their effects can be modelled and parameterized. This allows for using data from unstructured volunteers, by quantifying the spatial patterns of observer behaviour, to achieve better estimates of species richness, abundance, etc., without excluding large amounts of observations.

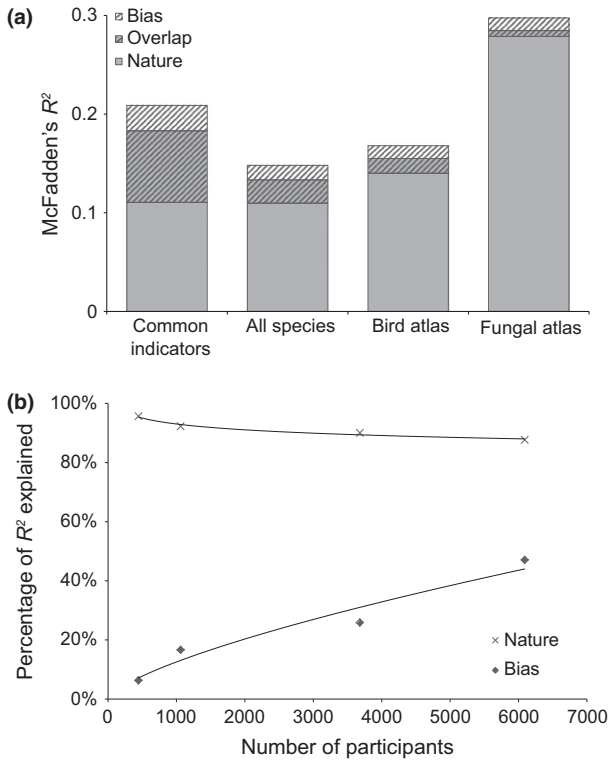

Figure 3 Variance partitioning based on McFadden's R^2 for the intensity of observation *for various contextual variables*. (a) The total hight of each bar is equivalent to the R^2 of the best fit model containing all variables. Grey bars show the R^2-values for the models containing only the land cover variable (nature). Crossed sections display the R^2 of the models containing roads and population density only (bias). The overlap is equivalent to the difference between the sum of the R^2 of the 'nature' model and the 'bias' model subtracting the R^2 of the full model. (b) shows the contribution of 'bias' (diamonds) and 'nature' (crosses) of the total R^2 as correlation with the number of participants in the individual recorder schemes.

Improving decision-making for conservation planning

An adequate and uniform sampling effort with sufficient spatial coverage and resolution is preferable when estimating species richness patterns (Woolhouse, 1983). However, this is often difficult to achieve in citizen science projects where data collection is based on volunteers (Boakes *et al.*, 2010). A few previous studies have shown that if indicators are selected carefully and if the amount of data is ample, even untrained volunteers can reach comparable or even more reliable species richness estimates than trained experts (Danielsen *et al.*, 2005; Goffredo *et al.*, 2010; Holt *et al.*, 2013), often at much larger spatial scales (Holt *et al.*, 2013). But most often, complete knowledge of all biodiversity values is not realistic or necessary to direct conservation action. Complete inventories may even skew the balance between resources spend on monitoring, and evaluation as opposed to resources allocated for conservation actions (Salzer & Salafsky, 2006). In this study, we identify areas of relative under- and oversampling based on land cover and infrastructure and combine these measures with estimates of species richness. We use this approach to map areas where the observed species richness is high or low compared to the intensity of sampling in the same area, correcting for spatial covariates. While such estimates should not be viewed as direct measures of true species richness, they may be used to guide conservation planning. In the absence of complete knowledge on species richness, the likelihood of an area being species rich given the available sampling effort could serve as a proxy for where to direct limited conservation resources. For example, in our study, sampling intensity of the *Fungal atlas* was strongly skewed towards forests. This could potentially pose a problem for species richness estimates as results could be an artefact of oversampling. However, accounting for spatial biases, our results show the opposite, with areas of Denmark that contain large amounts of forest having higher species richness than expected even after controlling for higher sampling effort (Figs 4a & S4). This would confirm that forests are good candidates for fungi biodiversity hotspots (Heilmann-Clausen *et al.*, 2015). In the other end of the spectrum, we see that agricultural landscapes have disproportionately low species richness even when accounting for the low number of observations (Figs 4a–c & S4). However, using expectations based on existing sampling structure can also enforce existing aggregated survey pattern where volunteer experts continue to survey areas expected to have higher biodiversity (Dennis *et al.*, 2006; Sastre & Lobo, 2009). In such cases, simpler volunteer schemes targeting all people have an advantage as they are less biased towards *a posteriori* expectations of where to search. Thus, mass-participating schemes such as our *All species* may be used to challenge presumptions about biodiversity that expert-driven datasets with limited coverage cannot.

Keeping to the point

Species records, whether collected by volunteers or professionals, can be described as the documentation of a biological phenomenon experienced at a given point. Such point observations often have a suite of information related to them, including exact time and place, observer ID and potentially metadata describing the purpose and extend of the observation effort. Such information is lost when points are aggregated into localities or a grid (Renner *et al.*, 2015). PPMs model how the observed intensity of points differs from a random point pattern. This has two overarching advantages, which are not possible in other models. First, it allows for modelling spatial covariables in a way that more closely resembles the expected underlying causal effect compared to a grid or locality-based approach. For example, we would expect the likelihood of observing a species to be a function of its distance to a road (Hanowski & Niemi, 1995; Keller & Scallan, 1999; Kadmon *et al.*, 2004; McCarthy *et al.*,

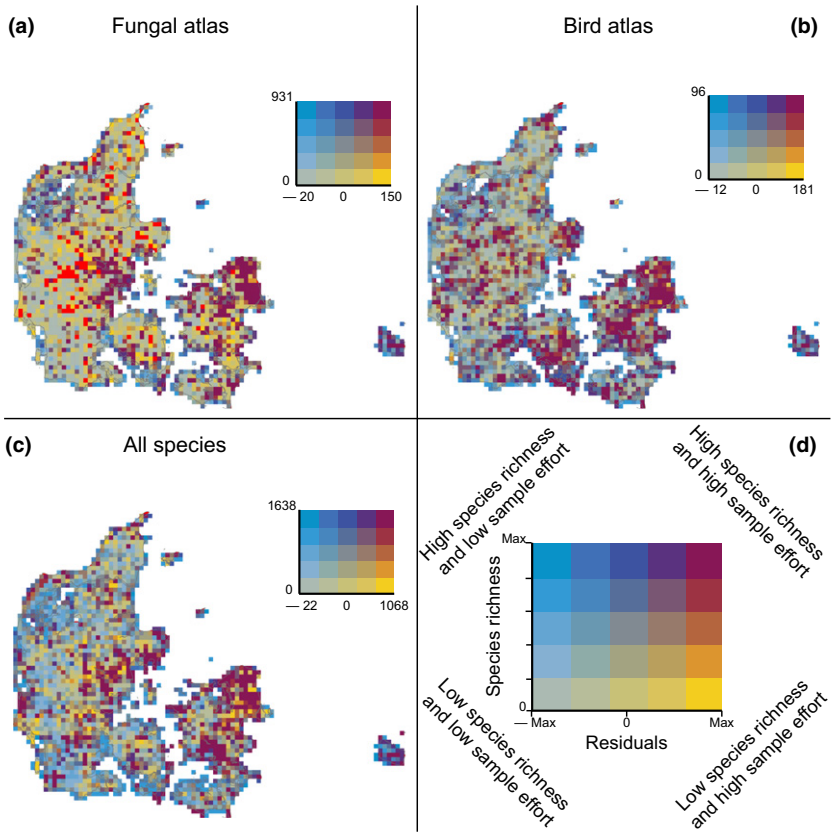

Figure 4 Bivariate map of residuals from the full model (*x*-axis) and species richness (*y*-axis). (a) Fungal atlas, (b) Bird atlas, (c) All species scheme and (d) Generic label. Blue areas in the top left corner are areas with high species richness and lower than expected sample efforts. These areas are expected to have a true higher species richness than expected. Yellow areas in the bottom right corner are areas with low species richness compared to their high sampling effort and the areas of true low species richness. Areas traversing diagonally from light bottom left to darker top right are areas where sample effort seems to match the observed species richness. Red areas are grid cells where residuals could not be calculated.

2012) or whether it is in a highly human populated area (Luck *et al.*, 2004; Luck, 2007). A grid- or polygon-based analysis reduces such an effect to a correlation between the number of observations and the total length of roads inside a grid cell, whereas PPMs allow us to model the effect in an exact spatial context. Likewise, the likelihood of observing a particular species is better described by whether an observation falls inside or outside an optimum habitat than the total area of that habitat within a given unit of analysis. A second advantage, which we do not explore in this study, is each point carries with it a suite of information which can be used to understand the intensity of observation. For example, weather can have an effect on both target taxa and volunteer behaviour both in time and space (Bas *et al.*, 2008), which can be easily captured by point-based analysis where such data can be attributed directly to the individual observation. Similarly, information linked to the species (e.g. red list status, size, daily rhythm) or the observer (e.g. skill level, number of records, sex) represents further important covariates (Kelling *et al.*, 2015). PPMs, unlike any other model approach, are able to address these factors, thus presenting a powerful approach where data on biases are available.

CONCLUSION

We show that volunteer-collected datasets that are more dependent on untrained amateurs are more heavily affected by spatial bias from infrastructure and human population density. Objectives and protocols of mass-participating citizen science projects should therefore be designed with this in mind. However, we also show that where data on spatial covariates are available, such data can be utilized to explicitly address biases. Thus, appropriately designed objectives, protocols and analyses of mass-participation schemes can produce results useful for management and research. Our results suggest that PPMs are a valuable amendment to a growing tool-case of advanced statistical tools for analysing unstructured volunteer-collected data. In areas with good contextual data, PPMs allow for understanding the effect of spatial bias at the level of individual observation, thus getting closer to the actual effect than classical grid-based analyses. Further, we suggest that PPMs allow for integrating more information related to the observation, which is currently often underutilized. Thus, while PPMs are currently not commonly used to analyse citizen science data, we see great potential in this approach for utilizing the enormous potential in volunteer-collected data.

ACKNOWLEDGEMENTS

First and foremost, we want to thank the thousands of volunteers who collected the data used in this analysis as well as the organizations and societies who helped set up the schemes. We would like to thank E. Rubak and A. Baddeley for invaluable guidance and assistance with point process models. Thanks to R.A. Holland for assistance with

illustrations. Financial support comes from Aage V. Jensen Naturfond, the Danish National Research Foundation (DNRF096), and VILLUM FONDEN (VKR023371).

REFERENCES

August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T. & Jepson, P. (2015) Emerging technologies for biological recording. *Biological Journal of the Linnean Society*, **115**, 731–749.

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Baddeley, A., Rubak, E. & Turner, R. (2015) *Spatial point patterns: methodology and applications with R*. Chapman and Hall/CRC, Boca Raton, FL, USA.

Barnes, M., Szabo, J.K., Morris, W.K. & Possingham, H. (2015) Evaluating protected area effectiveness using bird lists in the Australian Wet Tropics. *Diversity and Distributions*, **21**, 368–378.

Bas, Y., Devictor, V., Moussus, J.-P. & Jiguet, F. (2008) Accounting for weather and time-of-day parameters when analysing count data from monitoring programs. *Biodiversity and Conservation*, **17**, 3403–3416.

Bela, G., Peltola, T., Young, J.C., Balázs, B., Arpin, I., Pataki, G., Hauck, J., Kelemen, E., Kopperoinen, L., Van Herzele, A., Keune, H., Hecker, S., Suškevičs, M., Roy, H.E., Itkonen, P., Külvik, M., László, M., Basnou, C., Pino, J. & Bonn, A. (2016) Learning and the transformative potential of citizen science. *Conservation Biology*, doi:10.1111/cobi.12762.

Bell, S. *et al.* (2008) What counts? Volunteers and their organisations in the recording and monitoring of biodiversity. *Biodiversity and Conservation*, **17**, 3443–3454.

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

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference – a practical information-theoretic approach*, 2nd edn. Springer-Verlag, New York, USA.

Carvalheiro, L.G., Kunin, W.E., Keil, P., Aguirre-Gutiérrez, J., Ellis, W.N., Fox, R., Groom, Q., Hennekens, S., Van Landuyt, W., Maes, D., Van de Meutter, F., Michez, D., Rasmont, P., Ode, B., Potts, S.G., Reemer, M., Roberts, S.P.M., Schaminée, J., WallisDeVries, M.F. & Biesmeijer, J.C. (2013) Species richness declines and biotic homogenisation have slowed down for NW-European pollinators and plants. *Ecology Letters*, **16**, 870–878.

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J. & Waller, D.M. (2011) Assessing citizen science data quality: an invasive species case study. *Conservation Letters*, **4**, 433–442.

Danielsen, F., Burgess, N.D. & Balmford, A. (2005) Monitoring matters: examining the potential of locally-based approaches. *Biodiversity and Conservation*, **14**, 2507–2542.

Danmarks Statistik (2015) *Statistisk Årbog 2015*. Rosendahls A/S, Copenhagen, Denmark.

Dennis, R.L.H., Shreeve, T.G., Isaac, N.J.B., Roy, D.B., Hardy, P.B., Fox, R. & Asher, J. (2006) The effects of visual apparency on bias in butterfly recording and monitoring. *Biological Conservation*, **128**, 486–492.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, **10**, 291–297.

Ejrnæs, R., Wiberg-Larsen, P., Holm, T.E., Josefson, A., Strandberg, B., Nygaard, B., Andersen, L.W., Winding, A., Termansen, M., Hansen, M.D.D., Søndergaard, M., Hansen, A.S., Lundsteen, S., Baattrup-Pedersen, A., Kristensen, E., Krogh, P.H., Simonsen, V., Hasler, B. & Levin, G. (2011) *Danmarks biodiversitet 2010 – status, udvikling og trusler*, p. 152. Danmarks Miljøundersøgelser, Aarhus Universitet, Aarhus, Denmark.

Erb, P.L., McShea, W.J. & Guralnick, R.P. (2012) Anthropogenic influences on macro-level mammal occupancy in the Appalachian Trail corridor. *PLoS One*, **7**, e42574.

Eskildsen, A., Carvalheiro, L.G., Kissling, W.D., Biesmeijer, J.C., Schweiger, O. & Høye, T.T. (2015) Ecological specialization matters: long-term trends in butterfly species richness and assemblage composition depend on multiple functional traits. *Diversity and Distributions*, **21**, 792–802.

Gardiner, M.M., Allee, L.L., Brown, P.M.J., Losey, J.E., Roy, H.E. & Smyth, R.R. (2012) Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Frontiers in Ecology and the Environment*, **10**, 471–476.

Global Biodiversity Information Facility (2015) *GBIF*. Available at: http://www.gbif.org/occurrence (accessed 17 November 2015).

Goffredo, S., Pensa, F., Neri, P., Orlandi, A., Gagliardi, M.S., Velardi, A., Piccinetti, C. & Zaccanti, F. (2010) Unite research with what citizens do for fun: "recreational monitoring" of marine biodiversity. *Ecological Applications*, **20**, 2170–2187.

Gregory, R.D., van Strien, A., Vorisek, P., Gmelig Meyling, A.W., Noble, D.G., Foppen, R.P.B. & Gibbons, D.W. (2005) Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 269–288.

Hanowski, J.M. & Niemi, G.J. (1995) A comparison of on- and off-road bird counts: do you need to go off road to

count birds accurately? *Journal of Field Ornithology*, **66**, 469–483.

Heilmann-Clausen, J. & Læssøe, T. (2012) On species richness estimates, climate change and host shifts in wood-inhabiting fungi. *Fungal Ecology*, **5**, 641–646.

Heilmann-Clausen, J., Barron, E.S., Boddy, L., Dahlberg, A., Griffith, G.W., Nordén, J., Ovaskainen, O., Perini, C., Senn-Irlet, B. & Halme, P. (2015) A fungal perspective on conservation biology. *Conservation Biology*, **29**, 61–68.

Hickling, R., Roy, D.B., Hill, J.K., Fox, R. & Thomas, C.D. (2006) The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology*, **12**, 450–455.

Holt, B.G., Rioja-Nieto, R., Aaron MacNeil, M., Lupton, J. & Rahbek, C. (2013) Comparing diversity data collected using a protocol designed for volunteers with results from a professional alternative. *Methods in Ecology and Evolution*, **4**, 383–392.

Hörnsten, L. & Fredman, P. (2000) On the distance to recreational forests in Sweden. *Landscape and Urban Planning*, **51**, 1–10.

Isaac, N.J.B. & Pocock, M.J.O. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P. & Roy, D.B. (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5**, 1052–1060.

IUCN and UNEP-WCMC (2015) *The World Database on Protected Areas (WDPA)* [Nov 2015]. Available at: www.protectedplanet.net (accessed 2 November 2015).

Jepsen, M.R. & Levin, G. (2013) Semantically based reclassification of Danish land-use and land-cover information. *International Journal of Geographical Information Science*, **27**, 2375–2390.

Kadmon, R., Farber, O. & Danin, A. (2004) Effects of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Keller, C.M.E. & Scallan, J.T. (1999) Potential roadside biases due to habitat changes along breeding bird survey routes. *The Condor*, **101**, 50–57.

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F.A., Moore, T., Wiggins, A., Wong, W.-K., Wood, C. & Yu, J. (2015) Can observation skills of citizen scientists be estimated using species accumulation curves? *PLoS One*, **10**, e0139600.

Kéry, M. & Royle, J.A. (2016) *Applied hierarchical modeling in ecology – analysis of distribution, abundance and species richness in R and BUGS*. Academic Press, London, UK.

Kery, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Luck, G.W. (2007) A review of the relationships between human population density and biodiversity. *Biological Reviews*, **82**, 607–645.

Luck, G.W., Ricketts, T.H., Daily, G.C. & Imhoff, M. (2004) Alleviating spatial conflict between people and biodiversity. *Proceedings of the National Academy of Sciences USA*, **101**, 182–186.

Mackechnie, C., Maskell, L., Norton, L. & Roy, D. (2011) The role of 'Big Society' in monitoring the state of the natural environment. *Journal of Environmental Monitoring*, **13**, 2687–2691.

McCarthy, K.P., Fletcher, R.J. Jr, Rota, C.T. & Hutto, R.L. (2012) Predicting species distributions from samples collected along roadsides. *Conservation Biology*, **26**, 68–77.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. *Frontiers in econometrics* (ed. by P. Zarembka), pp. 105–142. Academic Press, New York.

Normander, B., Henriksen, C.I., Jensen, T.S., Sanderson, H., Henrichs, T., Larsen, L.E. & Pedersen, A.B. (2009) *Natur og Miljø 2009 – Del B: Fakta*. Danmarks Miljøundersøgelser, Aarhus Universitet, Aarhus, Denmark.

Peterson, A.T., Soberón, J., Pearson, R., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological niches and geographic distributions (MPB-49)*, p. 328. Princeton University Press, Princeton, NJ.

Powney, G.D. & Isaac, N.J.B. (2015) Beyond maps: a review of the applications of biological records. *Biological Journal of the Linnean Society*, **115**, 532–542.

Prendergast, J.R., Wood, S.N., Lawton, J.H. & Eversham, B.C. (1993) Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters*, **1**, 39–53.

Preston, C.D. (2013) Following the BSBI's lead: the influence of the Atlas of the British flora, 1962–2012. *New Journal of Botany*, **3**, 2–14.

R Development Core Team (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G. & Warton, D.I. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**, 366–379.

Roy, D.B., Harding, P.T., Preston, C.D. & Roy, H.E. (2014) *Celebrating 50 years of the Biological Records Centre*. Centre for Ecology & Hydrology, Wallingford, UK.

Salzer, D. & Salafsky, N. (2006) Allocating resources between taking action, assessing status, and measuring effectiveness of conservation actions. *Natural Areas Journal*, **26**, 310–316.

Sastre, P. & Lobo, J.M. (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Deri, E., Budrys, E., Kull, T., Tali, K., Bauch, B., Settele, J., Van Swaay, C., Kobler, A., Babij, V., Papastergiadou, E. & Henle, K. (2009) Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology*, **23**, 307–316.

van Strien, A.J., van Swaay, C.A.M. & Termaat, T. (2013) Opportunistic citizen science data of animal species

produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, **50**, 1450–1458.

Tewksbury, J.J., Anderson, J.G.T., Bakker, J.D., Billo, T.J., Dunwiddie, P.W., Groom, M.J., Hampton, S.E., Herman, S.G., Levey, D.J., Machnicki, N.J., del Rio, C.M., Power, M.E., Rowell, K., Salomon, A.K., Stacey, L., Trombulak, S.C. & Wheeler, T.A. (2014) Natural history's place in science and society. *BioScience*, **64**, 300–310.

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

Van Calster, H., Vandenberghe, R., Ruysen, M., Verheyen, K., Hermy, M. & Decocq, G. (2008) Unexpectedly high 20th century floristic losses in a rural landscape in northern France. *Journal of Ecology*, **96**, 927–936.

Warton, D.I. & Shepherd, L.C. (2010) Piosson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annual Applied Statistics*, **4**, 1–21.

Wiggins, A. & Crowston, K. (2011) From conservation to crowdsourcing: a typology of citizen science. *Proceedings of the 2011 44th Hawaii International Conference on System Sciences*, pp. 1–10. IEEE Computer Society, Kauai, HI, USA.

Woolhouse, M.E.J. (1983) The theory and practice of the species-area effect, applied to the breeding birds of British Woods. *Biological Conservation*, **27**, 315–332.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Supplementary figures and model validation.

Figure S1. Land cover map based on Jepsen & Levin (2013).

Figure S2. Residuals from the full models on number of observations.

Figure S3. Species richness for the four datasets.

Figure S4. Partial effects of roads for the four models.

Table S1. Land cover classes based on Jepsen & Levin (2013).

BIOSKETCH

The four citizen science data collecting schemes included in this study constitute the major efforts to involve volunteers in the collection of species records in Denmark. This study was developed to improve the use of citizen science data as well as build collaborations between biodiversity-focused citizen science projects in Denmark. The study brought together partners across the universities of Copenhagen and Aarhus as well as NGOs and volunteers involved in the setup and day-to-day running of the projects. The authors represent researchers in macroecology, ecology, taxonomy, nature conservation and statistics as well as many of the pioneers of Danish citizen science both inside and outside the universities.

Author contributions: J.G., J.H-C, C.R. and A.P.T. conceived the ideas; T.E.H, J.H-C., K.O., I.L., J.G., C.R. and A.P.T. coordinated the collection of data; data collections were performed by 1000s of volunteers; J.G., J.H-C, B.M. and A.P.T analysed the data; and all authors contributed to the writing.

Editor: Brian Leung