Received: 10 March 2020 | Accepted: 20 January 2021

DOI: 10.1002/2688-8319.12048

Ecological Solutions and Evidence | BRITISH ECOLOGICAL SOCIETY

CITIZEN SCIENCE

Research Article

Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection

Tanja K. Petersen[1,2] | James D. M. Speed[1] | Vidar Grøtan[2] | Gunnar Austrheim[1]

[1] Department of Natural History, NTNU University MuseumNorwegian University of Science and Technology, Trondheim, Norway

[2] Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Tanja K. Petersen, Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway.
Email: tanja.k.petersen@ntnu.no

Handling Editor: Ian Thornhill

Abstract

1. The availability and quantity of observational species occurrence records have greatly increased due to technological advancements and the rise of online portals, such as the Global Biodiversity Information Facility (GBIF), coalescing occurrence records from multiple datasets. It is well-established that such records are biased in time, space and taxonomy, but whether these datasets differ in relation to origin have not been assessed. If biases are specific to different types of datasets, and the relative contribution from these datasets have changed over time, these shifting biases will have implications for interpretations of results and, consequentially, for management and conservation measures.

2. We examined observational GBIF records from Norway to test potential differences in taxonomic, time and land-cover biases between 10 different datasets, with a focus on red-listed and non-native species.

3. The datasets differ in their taxonomic coverage, with datasets dominated by citizen scientist recorders focusing greatly on birds. The number of records has increased over time; in particular, citizen science datasets have had a sharp increase in recent years.

4. The different datasets (including division of the datasets by conservation status) showed differences in geographical coverage. Anthropogenic land covers have more records than would be expected by chance in the majority of cases. Remote areas have fewer records than would be expected, underlining the prevalence of a roadside bias.

5. Accounting for biases in opportunistic species occurrence records need to be a dynamic rather than static process, as the taxonomic and geographical biases have changed over time and differ between datasets, depending on origin and inherent characteristics. Data-collection programmes should be designed to counteract the biases of the specific datasets, and methods to account for the biases in existing data should be developed. When utilizing compiled, open-source data, care must be taken to ensure complementarity between the datasets, both regarding time and space. Incorporating strengths and accounting for biases between datasets can strengthen the integration

between species occurrence records with different origins for science-policy impact and management.

KEYWORDS
alien species, citizen science, Global Biodiversity Information Facility, land cover, museum collections, sampling bias, threatened species

1 INTRODUCTION

The amount and availability of data on species occurrences have increased tremendously in recent years (Gaiji et al., 2013), as have their use in applied conservation and biodiversity management (Powney & Isaac, 2015). Registering species occurrences have become far easier than in the early days of biogeographical surveys due to technological advancements and can be done with the help of volunteer amateurs ('citizen scientists') (Boakes et al., 2016). Online portals, for example the Global Biodiversity Information Facility (GBIF) (GBIF.org, 2019a), have further increased the public availability and interest (Amano et al., 2016). These portals gather data from various sources, ranging from digitized natural history collections to observations made by citizen scientists. Thus, these records are a mixture of data on preserved specimens and observational records from both structured surveys and opportunistic sightings (Speed et al., 2018). Volunteers participating in citizen science programmes (or autonomously reporting species occurrences) likely have different motivations for reporting than do institutional recorders registering species according to a specified aim, covering both intrinsic and extrinsic factors. For participants in citizen science programmes, the most important motivational factors have been reported as a personal connection, interest and concern for nature, a wish to contribute to science and (biodiversity and nature) conservation and the value/usefulness of their contributions (Ganzevoort, van den Born, Halffman, & Turnhout, 2017; Larson et al., 2020; Tiago, Gouveia, Capinha, Santos-Reis, & Pereira, 2017).

These mixed datasets thus suffer from various biases and errors due to their diverse origins and underlying motivations (Newbold, 2010). A frequently recognized bias for occurrence records is the 'roadside' bias; observations are reported more frequently short distances from roads and paths, due to easier accessibility (Kadmon, Farber, & Danin, 2004; Tye, McCleery, Fletcher, Greene, & Butryn, 2017). The term can be expanded to include areas near densely populated areas (Luck, 2007; Robinson, Ruiz-Gutierrez, & Fink, 2018). Concern has been raised repeatedly over this bias, especially if sampled areas cover significantly different environmental conditions than do un-sampled areas (Bystriakova, Peregrym, Erkens, Bezsmertna, & Schneider, 2012; Phillips et al., 2009; Speed et al., 2018). This potentially leads to faulty conclusions regarding biodiversity patterns (Kramer-Schadt et al., 2013). More importantly, if such potential biases are not similar among data providers (e.g. datasets mainly consisting of purely opportunistic citizen science records vs. datasets from structured, targeted institutional surveys), conclusions can differ depending on the proportional contribution from the different datatypes (Tye et al., 2017). Even more so, if this relative contribution from various types of datasets has changed over time.

In terms of biodiversity management, attention is frequently focused on specific taxonomic groups or on species of conservation concern (e.g. red-listed and alien species). However, different data providers might prioritize differently regarding taxonomic groups and species' management status (red-listed vs. alien). Citizen scientists can be biased towards charismatic, easily recognizable taxa (Amano et al., 2016) and have a greater incentive to report red-listed and rare species (Tulloch, Mustin, Possingham, Szabo, & Wilson, 2013). Speed et al. (2018) showed that observational plant records and preserved specimens have different biases regarding taxonomic coverage, time and space and hypothesized that these differences can be translated somewhat to whether the data originate from structured surveys or opportunistic records, thus illustrating some of the potential issues with these mixed datasets. Note, however, the distinction between observation- and specimen records is not equivalent to the distinction between citizen science- and institutional records; vegetation plot data will be registered as observations, and some specimens in natural history collections are supplied by citizens (Miller-Rushing, Primack, & Bonney, 2012; NTNU University Museum, 2018). Geldmann et al. (2016) showed that spatial bias in citizen science records depended on the sampling scheme, distance to roads and the human population density.

Understanding spatio-temporal dynamics of biodiversity is paramount to achieve sustainable management of biodiversity issues, for example red-listed and alien species; for example there is a general lack of understanding on how land use, a main but complex driver, affects biodiversity change, as detailed data on species occurrences associated with different land-use types often are limited. Fine-grain data on species distributions and associations from local to global spatial scales, and over long time periods are required – a task virtually impossible to achieve through targeted surveys alone (Bonney et al., 2009; Dickinson, Zuckerberg, & Bonter, 2010; Theobald et al., 2015). Opportunistic citizen science records are frequently used as a data source, for species distribution modelling (SDM) (Beck, Böller, Erhardt, & Schwanghart, 2014; Jetz, McPherson, & Guralnick, 2012), which can be used in decision-making for managing red-listed and alien species (Thuiller et al., 2005; Guisan et al., 2013; Syfert et al., 2014). As these models are sensitive to bias in the data (Yañez-Arenas, Guevara, Martínez-Meyer, Mandujano, & Lobo, 2014), methods to account for varying forms of bias in SDM's are still being explored (e.g. Kramer-Schadt et al., 2013; Dorazio, 2014; Robinson et al., 2018). A general caveat of using GBIF records in SDM is that only few of datasets report species absences, thus requiring the use of presence-only modelling.

If the inherent biases differ markedly between datasets collected through institutional surveys, as citizen science, or as a mixture of the two, and the proportional contribution from these groups has changed over time, this raises the additional issue of how to deal with shifting biases, rather than simply static ones.

To our knowledge, limited attention has been given to whether taxonomic, temporal and geographical sampling biases are similar for datasets with varying origins (i.e. predominantly from citizen science programmes or institutionally organized surveys), and whether these different datasets complement or amplify each other's biases. The same holds for records of conservation concern within these datasets (but see Beck et al., 2014, for a comparison of GBIF original source data; Tye et al., 2017, for comparison of SDMs based on citizen science or institutional observation records; Troudet, Grandcolas, Blin, Vignes-Lebbe, and Legendre, 2017, for an assessment of taxonomic bias in GBIF records; Speed et al., 2018, for comparison of spatial, environmental, temporal and taxonomic coverage of observational records vs. preserved specimens). Awareness of such differences can elucidate how such mixed datasets should be utilized in the future to ensure complementarity and what biases to account for. Specifically, it will provide guidance to (1) what geographical areas, taxonomic and conservation groups should be targeted to balance sampling effort (and by whom); (2) whether certain datasets (with specified origins and characteristics) are representative of all collected data, and if not: (3) how to ensure complementarity between datasets to obtain maximum coverage.

In this study, we aim to test the 10 datasets with the most records within the study region from GBIF, detailing their differences and biases in taxonomy, time and land-cover associations and relating these to the various origins and characteristics of the datasets. The datasets range from 'pure' opportunistic citizen science records to observations from structured, targeted surveys by scientific institutions. To relate the results to biodiversity management, focus will be put on red-listed and alien species.

We hypothesize the following:

H1: The distribution of records between the three main kingdoms (H1a) and alien- versus red-listed species (H1b) differ between the datasets; also within the datasets not explicitly focusing on a particular taxonomic group.

H2: There has been an increase in the number of records over time, primarily reflecting an increase in the activity of citizen scientists.

H3: The different datasets will be unevenly distributed among different land-cover types, with areas heavily influenced by humans (e.g. urban areas and agricultural land; areas classified as 'developed area' and 'cultivated' in Table S.1 in the Supporting Information (Figure 1, Figure S.1)) sampled more than would be expected by random chance; this oversampling is expected to be greater for datasets primarily consisting of citizen science records than for more targeted datasets.

2 | MATERIALS AND METHODS

2.1 | Land-cover and species occurrence records

The study was limited to Norway (Figure 1). This is a well-surveyed region regarding species occurrence records in GBIF (Chandler et al., 2017), covering great variation in land cover, climate, human population density and with detailed land-cover data available (Statistics Norway, 2020).

Land cover was based on the Norwegian AR50 maps from NIBIO (Norwegian Institute of Bioeconomy Research, 2019), downloaded through Geonorge (2019). Land cover is categorized based on land- and tree cover type, timber productivity and soil condition (Supporting Information S.1, Table S.1, Figure. S.1). Areas smaller than 1.5 ha are not visible in the dataset. The AR50 data were last updated in year 2016.

All georeferenced records of all taxa (regardless of taxonomic level) within Norway were downloaded from GBIF on 19 November 2019 (GBIF.org, 2019b). The full dataset consisted of 31,091,434 species occurrence records. Of these, 23,586,634 belonged to the kingdom Animalia, 1,275,533 belonged to Fungi, 5,872,214 belonged to Plantae, 283,924 belonged to Archaea, Bacteria, Chromista or Protozoa, 46 were viruses, and 73,083 had no reported kingdom or were *incertae sedis*. The records ranged temporally from 1686 to 2019.

The following criteria were used for improving the dataset quality and comparability: (1) records with the occurrence status '`absent`' were removed, as very few of the dominant datasets included information on absences. Thus, including absence records would reduce the comparability of the datasets; (2) records with no registered species-level information were removed to standardize the taxonomic resolution of the datasets; (3) potential duplicate records for species, date, basis of record, coordinates and coordinate uncertainty were removed, as there is no guarantee that the same records have not been registered multiple times by different data providers; (4) records from later than 31 December 2018 were removed, thus only including full sampling years. This was done in consideration of the temporal analyses; (5) only records classified as '`HUMAN_OBSERVATION`' were retained; as the distribution of data types differed greatly between datasets, only comparing data within a single basis of record increased the comparability among datasets. Only records from the kingdoms Animalia, Plantae and Fungi were retained. For the comparison of different datasets, the analyses were limited to datasets including >50,000 records. The final dataset for analyses consisted of 10 datasets holding a total of 7,560,590 records (Table 1; see Supporting Information S.2 and Table S.2 for detailed descriptions of the individual datasets). Most species were only observed sporadically (Supporting Information S.3 and Figures S.2 and S.3). The 10 datasets were not evenly distributed across Norway, neither individually nor in unison. However, as part of the aim of the study was to assess skews in geographical distribution, this was not considered an issue.

The datasets included in the analyses differ in origin and in several characteristics, including (but not limited to) taxonomic focus, methodology, number- and skill level of the reporters. Two of the datasets

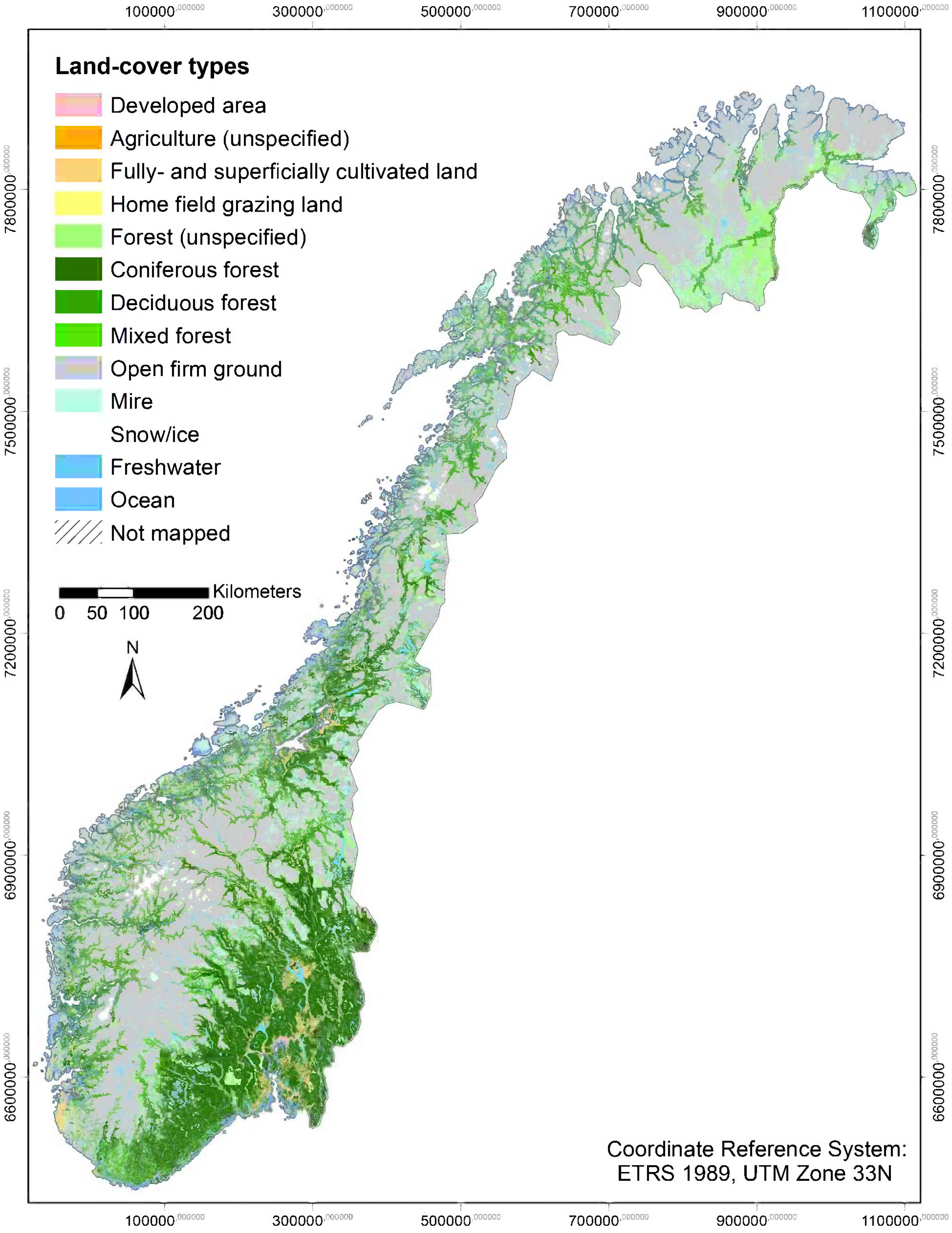

FIGURE 1 Map of Norway. Detailed maps of the individual land-cover types are shown in the Supporting Information S.1 and Figure S.1

TABLE 1 Datasets included in the analyses. The datasets are displayed in descending order according to the total number of records. The description is simplified from the description and methodology as presented on the GBIF webpage. More detailed descriptions can be found in Supporting Information S.2. and Table S.2

Dataset name (abbreviation)	Publisher (reference)	Number of animals Number of plants Number of fungi	Description
Norwegian Species Observation Service, (NBIC$_{CS}$)	Norwegian Biodiversity Information Centre (The Norwegian Biodiversity Information Centre and Hoem, 2020b)	2,678,373 1,776,878 422,930	Citizen science species observations (Artsobservasjoner.no)
Vascular Plants, Field notes, Oslo (UiO$_{Plant\ Notes}$)	Natural History Museum, University of Oslo (Natural History Museum, 2019a)	0 1,006, 937 0	Vascular Plants, Field notes from Norway
Norwegian Biodiversity Information Centre - Other datasets (NBIC$_{other}$)	Norwegian Biodiversity Information Centre (The Norwegian Biodiversity Information Centre and Hoem, 2020a)	585,898 17,288 289	Other data providers sharing occurrence data via the Norwegian Species Map Service
BioFokus (BioFokus)	BioFokus (Blindheim, 2020)	39,179 212,265 82,724	A non-profit organization providing biodiversity survey information on biological diversity to decision makers and the public in Norway.
Vascular plant field notes, NTNU University Museum (NTNU)	NTNU University Museum (NTNU University Museum, 2020)	0 188,910 0	Vascular plant field notes, Norway, using standardized cross-lists. The dataset is a collection of observations made during various research projects at the NTNU University Museum
Jordal (Jordal)	Biolog J.B. Jordal AS (Jordal, 2019)	282 150,411 21,856	John Bjarne Jordal, sole proprietor. Consultant within biology and nature management)
Vascular Plants, Field notes, Agder naturmuseum (KMN) (KMN)	Agder Museum of Natural History and Botanical Garden (Åsen, 2019)	0 125,115 0	Vascular Plants, Field notes from Norway
EOD – eBird Observation Dataset (eBird)	Cornell Lab of Ornithology (Levatich and Padilla, 2019)	93,109 0 0	eBird: a collective enterprise taking a novel approach to citizen science by developing cooperative partnerships among experts in a wide range of fields
Vascular Plants, Observations, Oslo (UiO$_{Plant\ Obs}$)	Natural History Museum, University of Oslo (Natural History Museum, 2019b)	0 82,634 0	Vascular Plants, Observations, Oslo
Lichen field notes, Oslo (UiO$_{Lichen}$)	Natural History Museum, University of Oslo (Natural History Museum, 2020)	0 0 75,512	Lichens, Field notes from Norway

can be regarded as 'pure' citizen science datasets ($\mathrm{NBIC_{CS}}$: Citizen science species observations from the Species Observation Service in Norway (The Norwegian Biodiversity Information Centre [NBIC] and Hoem, 2020b). eBird: citizen science records of birds, Levatich & Padilla, 2019)). Five datasets originated from museums and/or universities (KMN: vascular plant registrations from the Agder Museum of Natural History and Botanical Garden (Åsen, 2019); NTNU: vascular plant registrations according to standardized cross-lists (NTNU University Museum, 2020); $\mathrm{UiO_{Lichen}}$: lichen registrations from the University of Oslo Natural History Museum (Natural History Museum, 2020); $\mathrm{UiO_{Plant\ Obs}}$: vascular plant registrations (observational records) (Natural History Museum, 2019b); $\mathrm{UiO_{Plant\ Notes}}$: vascular plant registrations (field notes) (Natural History Museum, 2019a)) and can be regarded to cover somewhat structured surveys and observations by institutional recorders. Two datasets stemmed from a private consultant and organization (Jordal: consultant within biology and management (Jordal, 2019) and BioFokus: non-profit organization providing survey information (Blindheim, 2020)), which both provide biodiversity survey information for decision makers, and can thus be regarded as mainly structured surveys and observations done by institutional recorders. Likewise, the final dataset ($\mathrm{NBIC_{Other}}$) included datasets and databases from providers not hosting their own GBIF Integrated Publishing Toolkit (IPT) services, such as the Norwegian Environmental Agency – these are likewise regarded as mainly structured, institutional surveys. Data from NBIC are quality controlled internally: the data owner is responsible for the quality of the data. Dubious records are validated by experts, and the data owner is asked to provide evidence (e.g. photographs) of the record. If these cannot be provided, the record is deleted (Norwegian Biodiversity Information Centre [NBIC], 2018a; Norwegian Species Observation Service, 2020).

The latest Norwegian Red List of Species was finalized in 2015, 10 years after the first national assessment. The list includes species evaluated as being at risk of extinction in Norway, if conditions remain unchanged. The classification follows the same criteria as the IUCN Red List (Henriksen & Hilmo, 2015). In total, ≈4500 species are currently red-listed; of these are ≈2550 animals (mainly invertebrates), ≈750 plants and ≈1,200 fungi. The first version of the Alien Species List was compiled in 2007 (Gederaas, Moen, Skjelseth, & Larsen, 2012), and the latest version was refined and published in 2018 (Sandvik, Gederaas, & Hilmo, 2017; NBIC, 2018b). In total, ≈3000 species are listed as alien to mainland Norway, ≈1500 of these have a risk assessment. Of these, ≈390 are animals, ≈990 are plants and ≈100 are fungi. As per the guidelines published by the NBIC (Sandvik et al., 2017), we here use the term 'alien species' rather than the frequently used 'invasive species'. 'Alien' refers to '(…) a species introduced outside its natural past or present distribution' (IUCN, 2020). The term 'invasive' suggests invasion potential and negative ecological effects, which is not necessarily the case for all alien species. To avoid subjective decisions as to which alien species to classify as 'invasive', all species classified as 'alien' on the Alien Species List (Gederaas et al., 2012) were included, and the term 'alien' was used rather than 'invasive'.

Species names of the GBIF records were matched with the Norwegian Red List, and the Norwegian Alien Species List, using synonyms from the GBIF backbone taxonomy, using the package `rgbif` (Chamberlain & Boettiger, 2017). Species within the Red List categories 'regionally extinct', 'critically endangered', 'endangered', 'vulnerable', 'near threatened' and 'data deficient' are classified as 'red-listed'. As the majority of 'data-deficient' species are potentially threatened (Bland, Collen, Orme, & Bielby, 2015), and old records are included in the analyses, inclusion of the remaining Red List categories is warranted. Species alien to Svalbard, but native to mainland Norway were not listed as alien, neither were alien species which have not yet established, but are evaluated to have the potential to do so within 50 years; NBIC, 2018).

Maps and occurrence records were transformed to the geodetic coordinate reference system WGS84/UTM zone 32 (epsg: 32632).

2.2 | Statistical analyses

Taxonomic differences within and between datasets were examined using X^2- tests (base package: 'stats'), testing the null hypothesis of equal distribution of the kingdoms between and within the datasets. Likewise, the distribution of red-listed and alien species between the datasets was tested with a X^2-test.

To test for temporal patterns in the data, a Mann–Kendall test for a monotonic trend was applied (package: 'trend'; Pohlert, 2020). The median sampling year of the datasets were compared with a Kruskal–Wallis test, followed by a posthoc pairwise Dunn test with Bonferroni correction for multiple comparisons (packages: 'stats' and 'FSA'; Ogle, Wheeler, & Dinno, 2020).

For examining geographical biases, the data were further reduced to match the timeframe of the land-cover data. Only data from year 2004 to (and including) year 2018 were used. Changes in land cover are assumed to be minimal within this 15-year span. The remaining 5,622,260 records were overlaid on the AR50 map (package: 'sp'; Pebesma & Bivand, 2005). The null hypotheses was that the species occurrence records are randomly distributed across Norway, and the number of records is a function of the area of each land-cover type.

5,622,260 points were randomly overlaid on the map 100 times, giving ranges of expected number of points associated with each land-cover type. Dataset names and conservation status ('red-listed'-, and 'alien') were assigned randomly to the points in the same proportions as in the original data. Generalized linear models (Poisson error distribution, 'identity' link function) (base package: 'stats') were fitted to the number of records predicted by area of each land-cover type for the simulated data, providing the null models; one separate model for each of the combinations of dataset and conservation status. Sampling bias was concluded if the observed number fell outside the 0.95 confidence interval of the model. To compare the extent of sampling bias for the different groups, the absolute and relative residuals were calculated as

$$\begin{aligned}\text{Absolute residual} &= \text{Number of records}_{\text{observed}} \\ &\quad -\text{Number of records}_{\text{predicted}}\end{aligned}$$

and

$$\text{Relative residual} = \frac{\text{Absolute residual}}{\text{Mean}\left(\text{Number of records}_{\text{observed}}, \text{Number of records}_{\text{predicted}}\right)}.$$

To evaluate the differences in biodiversity patterns obtained using occurrence records from the different datasets, or all in combination, individual-based species accumulation curves were made for each dataset × conservation status group, and the asymptotic species richness calculated (package: ‘iNEXT’; Hsieh et al., 2020).

All data preparation and analyses were performed in R, version 3.6.1 (R Core Team, 2020). Maps were made in ArcMap version 10.6 (ESRI, 2018).

3 | RESULTS

3.1 | Taxonomic differences

The number of records from each dataset differed ($X^2 = 26{,}019{,}773$, $df = 9$, p-value < 0.001) with the vast majority of the records belonging to the NBIC_{CS} dataset, followed by the $\text{UiO}_{\text{Plant Notes}}$ (see Table 1 for description of dataset names). The kingdoms were not equally distributed between and within the datasets ($X^2 = 3{,}813{,}957$, $df = 18$, p-value < 0.001). Obviously, the datasets with a specified taxonomic scope were dominated by records belonging to the particular kingdom, but the datasets including several kingdoms differed as well; the BioFokus- and NBIC_{CS} datasets had an overabundance of animals and fungi, whereas the $\text{NBIC}_{\text{other}}$ dataset only had an overabundance of animal records. The Jordal dataset had an overabundance of plants and fungi (Figure 2). Within the animal kingdom, birds was the most frequently recorded class, followed by insects and mammals overall. For the multi-taxa datasets, the distribution within the animal kingdom differed: the BioFokus datasets held most records of insects, followed by birds and mammals, the Jordal dataset was dominated by birds, followed by insects and bivalves, and the NBIC_{CS}- and $\text{NBIC}_{\text{Other}}$ datasets were dominated by records of birds, followed by insects and mammals (Figure S.4 in the Supporting Information). When accounting for the different sample sizes, the distribution of red-listed and alien species differed between the datasets, with the BioFokus, eBird, NBIC_{CS}, $\text{NBIC}_{\text{other}}$ and $\text{UiO}_{\text{Lichen}}$ holding more red-listed, and the KMN, Jordal, NTNU, $\text{UiO}_{\text{Plant Nores}}$ and $\text{UiO}_{\text{Plant Obs}}$ datasets holding more alien species than what would be expected by random ($X^2 = 104{,}807$, $df = 9$, p-value < 0.001; Figure 2(b)).

3.2 | Temporal differences

The Mann–Kendall test detected a tendency in the overall dataset; the number of records had increased over time ($z = 16.732$, $n = 200$, p-value < 0.001; Figure 3(a)). Median year differed for all datasets (medians: KMN = 1986, BioFokus = 2011, eBird = 2015, Jordal = 2007, $\text{NBIC}_{\text{CS}} = 2014$, $\text{NBIC}_{\text{other}} = 2014$, NTNU = 1985, $\text{UiO}_{\text{Lichen}} = 2000$, $\text{UiO}_{\text{Plant Notes}} = 1961$, $\text{UiO}_{\text{Plant Obs}} = 2009$, Kruskal–Wallis = 496.44, $df = 9$, p-value = < 0.001. p-value < 0.001 for all pairwise comparisons; Figure 3(b)).

3.3 | Geographic differences

The simulated numbers of records within the groups (conservation status × dataset) were predicted by the area of the specified land cover type (Table 2, Figure 4).

Each land-cover type was relatively over- or under-sampled for different datasets (the observed number of records fell outside of the 0.95 confidence interval of models based on the simulated data), except for snow/ice, which was under-sampled by all datasets. The results are summarized in Table 3, and the full table can be seen in the Supporting Information S.6.

Models and results regarding datasets (regardless of conservation status) can be seen in the Supporting Information (Supporting Information S.5).

Comparing the absolute residuals between predicted and observed number of records within each land-cover type, the largest numerical discrepancies were seen for open firm ground, developed areas and cultivated land (Figure 5(a)). However, comparing the relative residuals (disregarding un-mapped areas and snow/ice), only alien records associated with open firm ground showed a consistent pattern between datasets (under-sampling; Figure 5(b)).

3.4 | Asymptotic species richness

The asymptotic species richness differed for most of the datasets (Supporting Information S.7). For both red-listed- and alien species, only the estimates for the NBIC_{CS} datasets (NBIC_{CS} red-listed = 2,412 [CI = 2 333–2 513], NBIC_{CS} alien = 867 [CI = 833-920]) overlapped with the estimates for all datasets combined (combined red-listed = 2 550 [CI = 2 469–2 654], combined alien = 861 [CI = 836–902]).

4 | DISCUSSION

Various forms of biases have been shown for the increasing amount of species data available from open databases, such as GBIF. However the potential taxonomic, temporal and geographical biases differ between datasets according to the origin and characteristics of the datasets, and how these different datasets might complement each other, have not been addressed. Additionally, whether these biases extend to red-listed and alien species remain un-investigated. We found that multi-taxa datasets from GBIF are biased towards different kingdoms (supporting H1a). More records of red-listed species are registered than alien species; (supporting H1b). When categorizing the records according to datasets and conservation status, the geographical biases

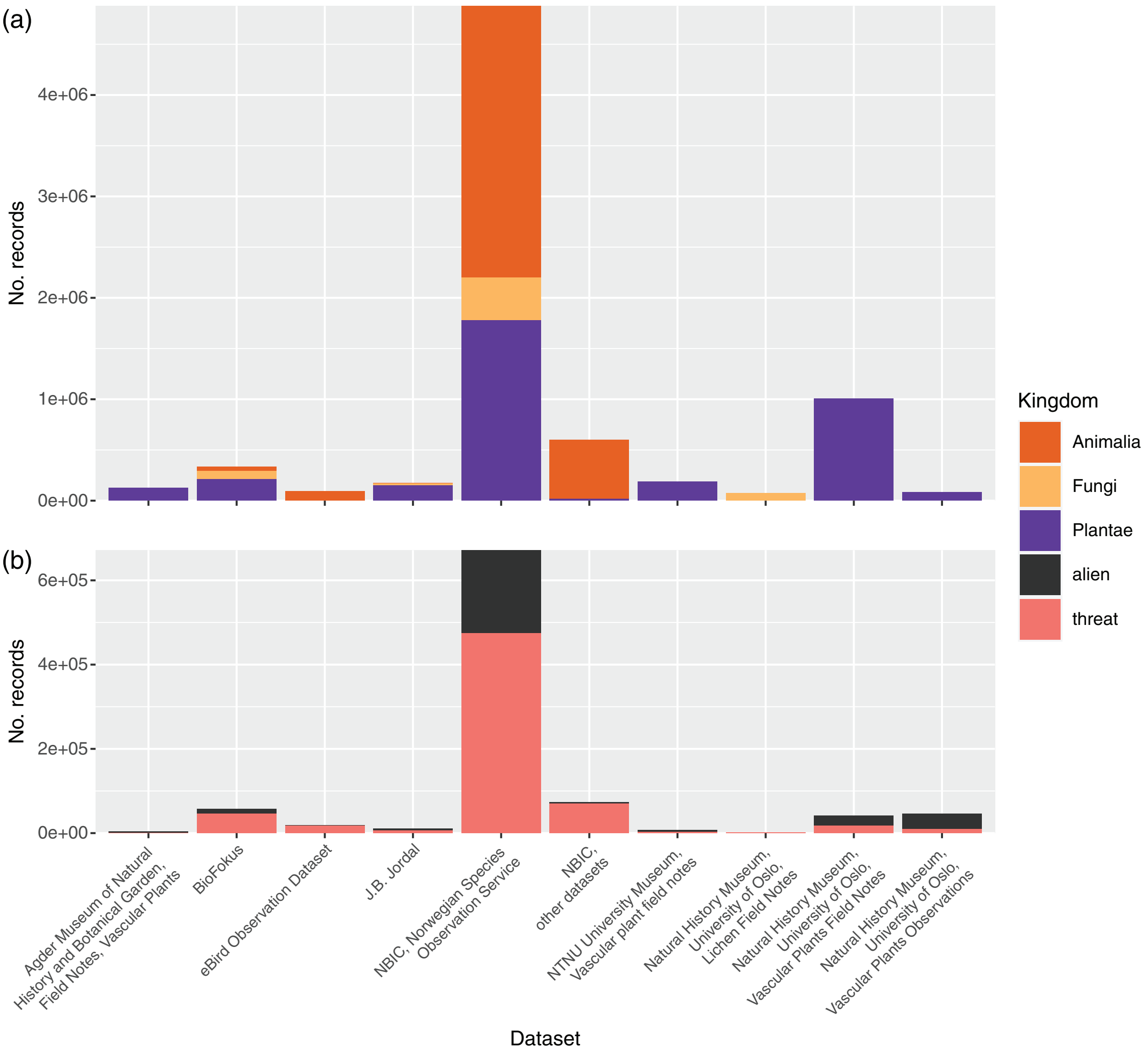

FIGURE 2 Number of records within each of the datasets used in the analyses. (a) Number of records from the included kingdoms in each dataset; (b) number of red-listed- or alien species records in each dataset. Note the differences in *y*-axis values due to species neither on the Red List nor the Alien Species List included in (b)

differ between the datasets, with a few general patterns. Anthropogenic land covers are generally oversampled (with a few exceptions), whereas less directly human-affected- and/or remote areas are undersampled (somewhat supporting H3).

4.1 | Differences in taxonomic groups and conservation status between datasets

The taxonomic bias within and between the datasets differ markedly, both in the sense that several of the datasets are concerned with a single taxonomic group, and in that the multi-taxa datasets are skewed towards a single group. The datasets originating from museums all focus on plants (except for $\text{UiO}_{\text{Lichen}}$; lichens are here classified as fungi). These patterns are reflected when comparing the multi-taxa datasets: the two datasets from the NBIC are both dominated by animal records, whereas the BioFokus and Jordal are both dominated by plants. Interestingly, only two out of the 10 datasets can be regarded as citizen science, but yet they make up the bulk of the records. The dominance of birds within these datasets reflect the long-term popularity of ornithology (Devictor, Whittaker, & Beltrame, 2010), the incentive for people to report on charismatic, recognizable species, and that many citizen science programmes have focused on birds (Tulloch et al., 2013). However, if the datasets dominated by citizen science records are not considered, the avian dominance is much less pronounced. This echoes the taxonomic differences observed by Troudet et al. (2017) and Speed et al. (2018). Theobald et al. (2015) found the taxonomic bias in citizen science and institutional datasets to be consistent; however, they did see an overweight of respectively birds and plants in the two groups. This underlines the careful considerations which much be taken eventually when using citizen science in multi-taxa analyses – nevertheless, within popular taxa, citizen science records can be a useful supplement

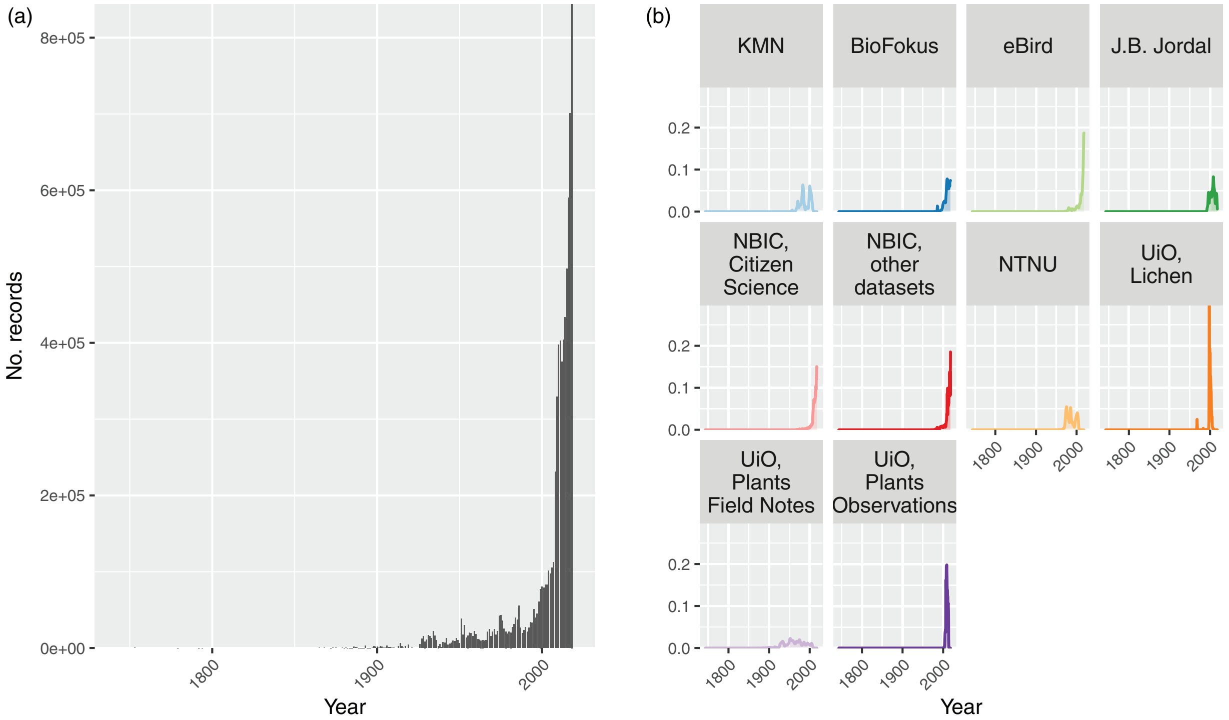

FIGURE 3 (a) Number of GBIF records across years in total. (b) Density plots of the number of records, divided by datasets. Note that the *y*-axis in (b) indicate proportion rather than absolute number. Acronyms refers to the datasets described in Table 1

to institutional observations, as this allows for otherwise impossible sample sizes (Tulloch, Possingham, et al., 2013; Powney & Isaac, 2015). Citizen science data on popular taxa have proven useful for discovering population trends, conservation and management (e.g. for birds: Lehikoinen et al., 2019; and examples in Sullivan et al., 2009).

The datasets with more alien- than red-listed records were all datasets focused on vascular plants; for all other datasets, more red-listed- than alien records were registered. This illustrates that most species on the Alien Species List are plants (NBIC, 2018). The dominance of red-listed species compared to alien species among half of the datasets, in particular in the datasets dominated by citizen science records (NBIC$_{CS}$ and eBird) points to a greater interest for rarities among citizen scientists and a potential lack of interest or knowledge regarding alien species. Among the other datasets, the difference can be due to a traditionally larger focus on red-listed species, or that red-listed species are likely registered as observations (i.e. not destructively sampled; NTNU University Museum, 2018), whereas alien species are potentially sampled as specimens to ensure validation later. The numerical difference between the conservation status groups can nevertheless be an artefact of the number of species in either status group: approximately three times as many species are on the Norwegian Red List compared to the Alien Species List (NBIC, 2015, 2018).

4.2 | Geographical biases

The most anthropogenic land-cover types have higher numbers of records than what would be expected for most, but not all groups. Developed areas were oversampled overall in all but three datasets (KMN, NTNU and Jordal); when focusing on either red-listed or alien records, the same pattern emerges, with the exception of the Jordal dataset being oversampled and the UiO$_{\text{Plant Obs}}$ being undersampled for red-listed species. This pattern likely has multiple underlying causes: despite a general omission of cities in ecological history (reviewed by Salomon Cavin & Kull, 2017), the last decades have seen increased focus on urban ecology, especially on cities as centres of spread for alien species (Gaertner et al., 2017). This has likely amplified the oversampling of alien species in urban areas. The oversampling of red-listed species is likely a combined effect of roadside bias and interest/prestige, as the oversampling is particularly large for datasets dominated by citizen science records.

Agricultural areas are similarly oversampled for most groups. This again reflects the roadside bias, as agricultural areas are generally found near developed areas (Figure S.1 in the Supporting Information), and thus have high accessibility. Grazing land is particularly oversampled, reflecting how such areas are regarded as of conservation concern, thus warranting attention from different recorders (Pärtel, Bruun, & Sammul, 2005).

The picture is highly nuanced regarding the different forest categories. The cases of oversampling may reflect that sampling tends to be done where high species richness is expected a priori (Boakes et al., 2016), the high amount of woodland in Norway (>35%), and the high species richness of forests (≈60% of Norwegian species are associated with woodlands). The highest number and concentration of red-listed species is found in coniferous woodlands and broad-leaved deciduous woodland, respectively (Gjerde, Brandrud, Ohlson, & Ødegaard, 2010),

TABLE 2 Model output. Simulated occurrence data randomly distributed across the AR50 map; conservation status and dataset name assigned in the same proportions as for the GBIF data (100 repetitions). Generalized linear models (Poisson error distribution, '`identity`'-link function) of the simulated data were fitted, predicting number of records falling within each land cover by the area of the respective land-cover type. P-values below 0.05 are highlighted in bold text. Acronyms refers to the datasets described in Table 1

(a) Red-listed species occurrence records				
	Estimate	**Standard error**	***z*-value**	***p*-value**
Dataset: KMN				
Intercept	3.337e-04	3.618e-06	−92.25	<0.001
Proportion of total area	9.218e+01	9.601e-01	96.00	<0.001
Dataset: BioFokus				
Intercept	−4.525e-02	3.253e-02	−1.39	0.164
Proportion of total area	4.223e+04	2.056e+01	2054.50e	<0.001
Dataset: eBird				
Intercept	−2.288e-03	2.173e-02	−0.105	0.916
Proportion of total area	1.397e+04	1.182e+01	1181.673	<0.001
Dataset: Jordal				
Intercept	1.002e-02	1.681e-02	0.596	0.551
Proportion of total area	5.324e+03	7.300e+00	729.297	<0.001
Dataset: NBIC$_{CS}$				
Intercept	−−8.217e-02	1.181e-01	0.696	0.486
Proportion of total area	4.165e+05	6.456e+01	645.341	<0.001
Dataset: NBIC$_{Other}$				
Intercept	−1.640e-03	4.370e-02	−0038	0.97
Proportion of total area	5.451e+04	2.335e+01	2333.829	<0.001
Dataset: NTNU				
Intercept	−3.694e-04	3.680e-06	−100.4	<0.001
Proportion of total area	1.020e+02	1.010e+00	101.0	<0.001
Dataset: UiO$_{Lichen}$				
Intercept	−9.760e-04	6.637e-06	−147.1	<0.001
Proportion of total area	2.696e+02	1.642e+00	164.2	<0.001
Dataset: UiO$_{Plant\ Notes}$				
Intercept	−2.406e-03	2.899e-05	−83.01	<0.001
Proportion of total area	6.647e+02	2.578e+00	257.81	<0.001
Dataset: UiO$_{Plant\ Obs}$				
Intercept	2.763e-02	2.474e-02	1.117	0.264
Proportion of total area	9.981e+03	9.996e+00	998,450	<0.001

(b) Alien species occurrence records				
	Estimate	**Std. error**	**z-value**	**p-value**
Dataset: KMN				
Intercept	−1.310e-03	7.510e-05	−17.45	<0.001
Proportion of total area	3.620e+02	1.903e+00	190.27	<0.001
Dataset: BioFokus				
Intercept	4.047e-02	2.651e-02	1.527	0.127
Proportion of total area	9.240e+03	9.619e+00	960.567	<0.001
Dataset: eBird				
Intercept	1.471e-02	1.137e-02	1.294	0.196
Proportion of total area	3.658e+02	1.919	190.657	<0.001

(Continues)

TABLE 2 (Continued)

(b) Alien species occurrence records	Estimate	Std. error	z-value	p-value
Dataset: Jordal				
Intercept	2.351e-02	1.726e-02	1.362	0.173
Proportion of total area	2.442e03	4.948e+00	493.657	<0.001
Dataset: NBIC$_{CS}$				
Intercept	5.979e-04	8.174e-02	0.007	0.994
Proportion of total area	1.889+05	4.347e+01	4344.328	<0.001
Dataset: NBIC$_{Other}$				
Intercept	8.834e-03	1.390e-02	0636	0.525
Proportion of total area	3.120e+03	5.598e+00	558.283	<0.001
Dataset: NTNU				
Intercept	−3.128e-04	2.901e-05	−10.78	<0.001
Proportion of total area	8.640e+01	9.296e-01	92.95	<0.001
Dataset: UiO$_{Plant\ Notes}$				
Intercept	−5.791e-03	2.618e-05	−221.2	<0.001
Proportion of total area	1.600e+03	4.000e+00	399.9	<0.001
Dataset: UiO$_{Plant\ Obs}$				
Intercept	1.108e-02	3.710e-02	0.299	0.765
Proportion of total area	3.595e+04	1.897e+01	1895.303	<0.001

which is somewhat seen in the positive residuals of red-listed records from most datasets. Some of the datasets hold fewer red-listed records than expected for coniferous- (KMN, eBird, Jordal NBIC$_{\mathrm{CS}}$ (red-listed), and UiO$_{\mathrm{Plant\ Obs}}$) and deciduous (eBird, NBIC$_{\mathrm{CS}}$ (red-listed), NTNU (red-listed), UiO$_{\mathrm{Lichen}}$, and UiO$_{\mathrm{Plant\ Obs}}$) forests. This discrepancy presumably stems from the taxonomical difference between the datasets: red-listed woodland species in Norway are mainly fungi, insects and lichens (Gjerde et al., 2010; Henriksen and Hilmo, 2015), and the number of red-listed plants outnumber red-listed animals; according to the NBIC (2015), only 14 out of 82 red-listed birds are associated with forests. Both datasets mainly collected by citizen scientists are heavily dominated by (or exclusively consists of) birds, which are easier to observe in open areas. Unclassified forests have fewer records than predicted for almost all datasets, except NBIC$_{\mathrm{Other}}$ and UiO$_{\mathrm{Plant\ Obs}}$, reflecting that this forest type is found in more remote, inaccessible areas; these two datasets have likely targeted such areas specifically. The land covers with fewer records than predicted for most of the datasets are characterized by being located in more remote and/or inaccessible areas, less directly affected by humans: snow/ice-covered areas, mires and open firm ground. In some instances, this reflects genuine low species richness and abundance (discussed below), as is likely the case for 'snow/ice' (having the largest relative residuals) and the most alpine cases of 'open firm ground'. However, some areas are likely under-sampled due to inaccessibility (e.g. mires), genuine difference in spatial- and taxonomic focus and interest of the datasets.

The discrepancies between predicted and observed number of records should be interpreted with caution. Some land-cover types are naturally more species poor than others. It can thus be expected that a lower number of records should be reported, than would be expected solely from area. This is the case of alpine areas; it is estimated that only ≈14% of the native vascular plants of Norway occur in mountains (Austrheim, Bråthen, Ims, Mysterud, & Ødegaard, 2010). Alpine areas are here found within the land-cover types 'open firm ground' and 'snow/ice', both of which have fewer records than predicted by the null models. Consequently, parts of the differences between observations and predictions can be attributed to the null models not taking intrinsic differences in species richness and abundances into account. Nevertheless, as we were not modelling species richness, but number of records (a proxy of sampling effort), the main signals are mainly attributable to differences in sampling effort.

4.3 | Dataset complementarity

The general quality of the data found in open databases, such as GBIF, is a point worth general discussion. Various opinions on the matter exist (Gaiji et al., 2013; Newbold, 2010; Powney & Isaac, 2015). The biases shown underlines how the individual datasets stored in GBIF are not all compiled with the intention of covering all taxa, periods or habitat. Thus, indiscriminately using such compiled datasets without accounting for the differences in sampling effort (whether this is spatial, temporal or taxonomic) will inevitably lead to flawed results. The differences in both taxonomic and geographic focus of different datasets from open databases shown in this study raise the question on

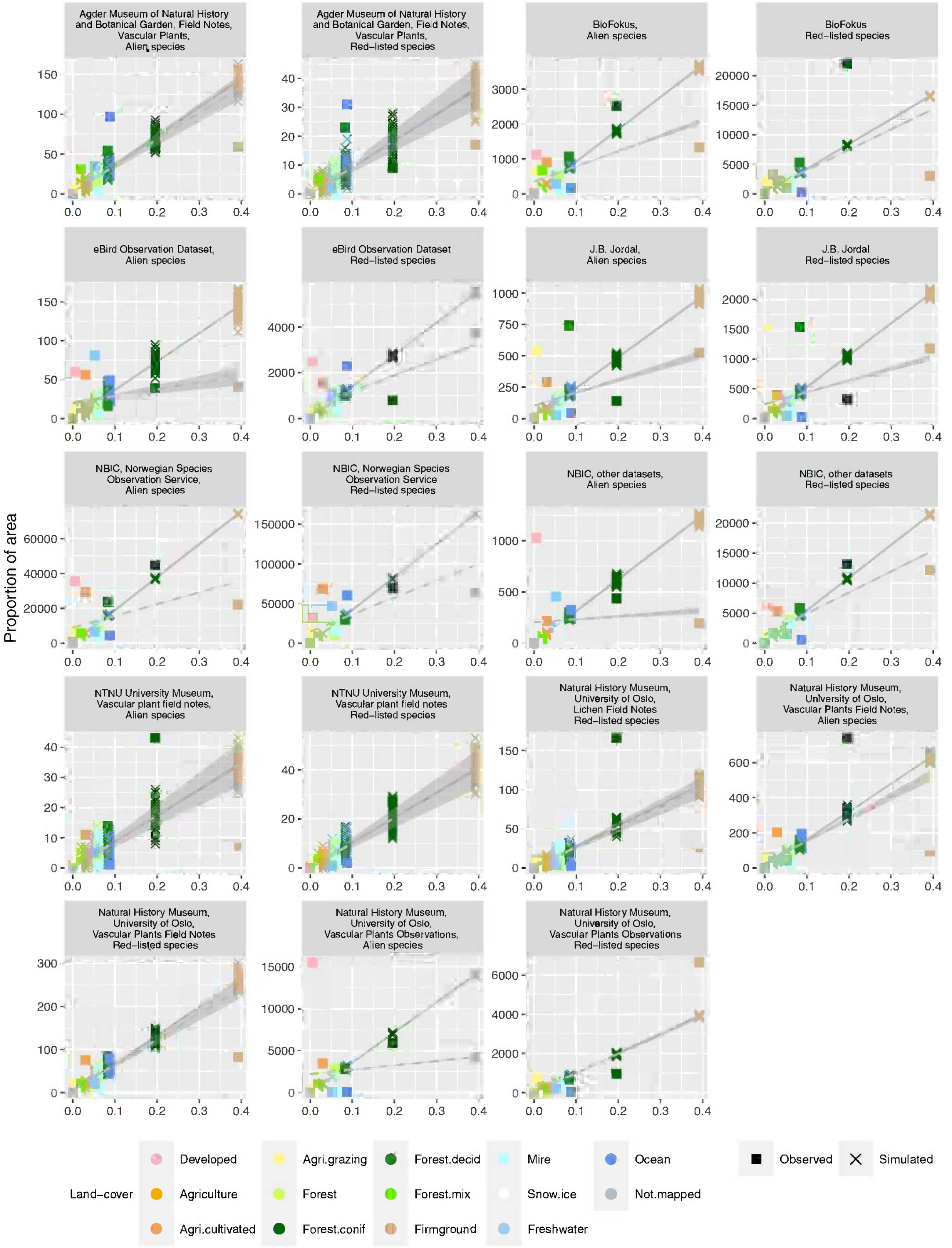

FIGURE 4 Null models (GLM) of the number of records as a function of area (proportion of total area within Norway) for the simulated data (crosses), vs. the observed number of records for each land-cover type (squares). Solid lines indicate model predictions; grey ribbons indicate the 0.95 confidence interval. Dashed lines indicate regressions similar to the null-models fitted through the observed values.

TABLE 3 Over- versus under-sampled land-cover types for each dataset. A summary of which land-cover types has either more or fewer observed records than expected by the Generalized Linear Models summarized in Table 2. ↑ indicates more records than expected, ↓ indicates fewer records than expected. 'n.s.' indicates that the observed number of records fell within the 0.95 C.I. of the model predictions. (See Supporting Information, Table S.6 for detailed numbers.) Acronyms refers to the datasets described in Table 1

	KMN	BioFokus	eBird	Jordal	NBIC$_{\text{CS}}$	NBIC$_{\text{Other}}$	NTNU	UiO$_{\text{Lichen}}$	UiO$_{\text{Plant Notes}}$	UiO$_{\text{Plant Obs}}$
	Red-listed	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*	*Red-listed*
	Alien	*Alien*	*Alien*	*Alien*	*Alien*	*Alien*	*Alien*	–	*Alien*	*Alien*
Developed area	↓	↑	↑	↑	↑	↑	↓	↑	↑	↓
	↓	↑	↑	↓	↑	↑	↓	–	↑	↑
Agriculture (unsp.)	↓	↑	↓	↓	↑	↑	↓	↓	↓	↓
	↓	↑	↓	↓	↑	↑	↓	–	↓	↑
Cultivated land	↑	↑	↑	↑	↑	↑	↑	↓	↑	↓
	↓	↑	↑	↑	↑	↑	↑	–	↑	↑
Home fields grazing land	↓	↑	↑	↑	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑	↓	–	↑	↑
Forest (unsp.)	↓	↓	↓	↓	↓	↑	↓	↓	↓	↓
	↓	↓	↓	↓	↓	↓	↓	–	↓	↑
Coniferous forest	↓	↑	↓	↓	↓	↑	↑	↑	↓	↓
	↓	↑	↓	↓	↑	↓	↑	–	↓	↓
Deciduous forest	↑	↑	↓	↑	↓	↑	↓	↓	↑	↓
	↑	↑	↓	↑	↑	↓	↑	–	↓	↓
Mixed forest	↑	↑	↓	↑	↑	↑	↑	↓	↑	↑
	↑	↑	↓	↓	↑	↓	↑	–	↑	↑
Open firm ground	↓	↓	↓	↓	↓	↓	↑	↓	↓	↓
	↓	↓	↓	↓	↓	↓	↓	–	↓	↑
Mire	↓	↓	↓	↓	↓	↓	↓	↑	↓	↓
	↓	↓	↓	↓	↓	↓	↓	–	↓	↓
Snow/ice	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	↓	↓	↓	↓	↓	↓	↓	–	↓	↓
Freshwater	↓	↓	↑	↓	↑	↓	↓	↓	↓	↓
	↑	↓	↑	↓	↓	↑	↑	–	↓	↓
Ocean	↑	↓	↑	↓	↑	↓	↓	↓	↑	↓
	↑	↓	↑	↓	↑	↑	↓	–	↑	↓
Not mapped	n.s.	↑	↑	n.s.	↓	↑	n.s.	n.s.	n.s.	↓
	n.s.	↑	n.s.	n.s.	↑	n.s.	n.s.	–	n.s.	↓

how to compile such datasets to ensure optimal coverage, and whether datasets with certain origin and characteristics are complementary. If multi-taxa management decisions are to be made based on analyses including, for example GBIF data, several considerations must be taken into account:

1. Regarding taxonomic complementarity, it is clear that careful examination of the included datasets is necessary, as indiscriminate data use will result in taxonomic imbalances.
2. Likewise, as the temporal coverage of the datasets is highly variable, timespan of individual datasets should be considered in relation to the questions asked.
3. Considering the geographical dissimilarities between the datasets, it is evident that if conclusions regarding the importance of different land-cover types for species of conservation concern are drawn upon analyses of single datasets, contrasting results will follow.
4. The geographic coverage of the single datasets used in analyses should be investigated to ensure that certain areas are nor over- or under-represented.

The overarching theme of these points can be summarized as not to assume a greater quality and validity of the available data than what is warranted. Care must be taken as to not stretch the conclusions based

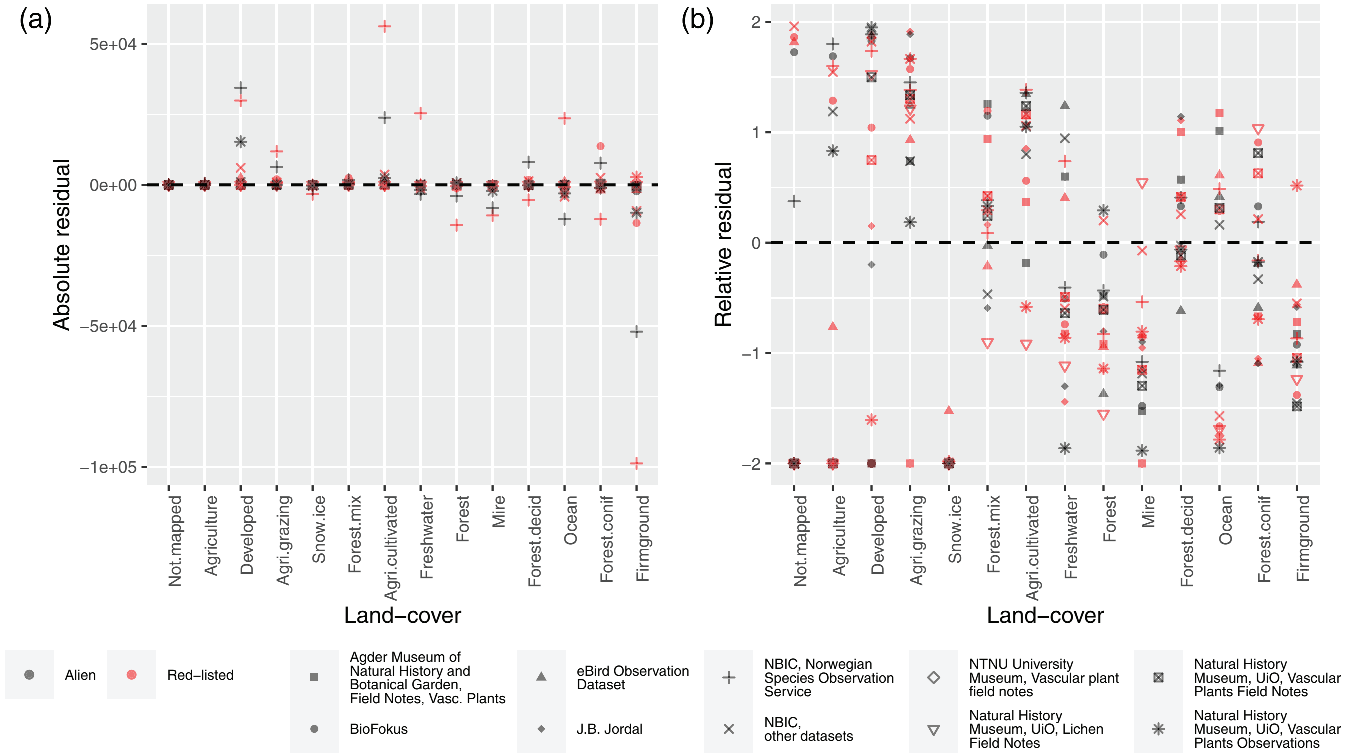

FIGURE 5 Differences between observed number of records within each land-cover type and the number of records predicted by area. (a) Absolute residuals (Number of $\text{records}_{\text{observed}}$ − Number of $\text{records}_{\text{predicted}}$); (b) relative residuals ($\frac{\text{Absolute residual}}{\text{mean(Number ofrecords}_{\text{observed}}\text{, number of records}_{\text{predicted}}\text{)}}$). Colours indicate conservation status, shapes indicate dataset. The land-cover types are ordered increasingly with respect to area

on single datasets further than the extents of the individual datasets, geographically or taxonomically.

4.4 | Integrating multiple datasets for understanding and managing biodiversity

Data availability thus remains the main challenge for understanding biodiversity patterns, and ultimately for how we manage biodiversity (Magurran, Dornelas, Moyes, & Henderson, 2019). This study has examined how different datasets, with different origins and characteristics, can complement each other in filling data availability gaps, specifically the gaps for three kingdoms (animals, plants and fungi), red-listed and alien species and their distributions across land covers and time.

Despite the emerging paradigm of data reuse and sharing among scientists, lack of data publishing is still an issue; only 10% of bio-collections is estimated to be digitally available, including data used prior to recent changes in data publishing policies provided by funding agencies and journals (Ball-Damerow et al., 2019). Traditionally, most collected data have been stored locally, and data not directly used in publications have remained unused and potentially forgotten with time (Osawa, 2019). This also leaves the worst case scenario that not all parts of datasets are published. Likewise, standardization of biodiversity data among data providers is important to ensure interoperability (Poisot, Bruneau, Gonzalez, Gravel, & Peres-Neto, 2019). An attempt at this is to use the Darwin Core Archive format adopted by GBIF (Osawa, 2019; Wieczorek et al., 2012). Despite these efforts, substantial quantities of primary biodiversity data (and metadata) remain undiscovered (Chavan & Penev, 2011). This leaves a gap in the foundation of biodiversity research. In the light of the results presented here, if the lack of data sharing is uneven among datasets with different origins, the gap is even more severe.

Open source, compiled biodiversity data have potential to be used for biodiversity modelling, if spatially biased sampling effort can be corrected for (Higa et al., 2015). Unfortunately, a recent review found that only 69% of the examined papers addressed some aspect of data quality (Ball-Damerow et al., 2019). Our results caution that careful considerations of the data used in such studies are needed; as the contribution from different datasets have changed over time, so has the geographical bias. Therefore, accounting for bias should be a dynamic process, dependent on timespan of the included data and the data contributors. If observational datasets of mixed origins are used indiscriminately, the reported spatio-temporal patterns could merely reflect spatio-temporal shifts in bias. Future surveys and citizen science programmes should aim to include otherwise neglected taxonomic groups, especially in under-sampled land-cover types, such as remote mountainous areas. In particular, non-avian animals are underrepresented compared to their actual abundance, and open firm ground and mires should be investigated more closely. Citizen science programmes focusing on non-avian taxa should be designed, learning from the success of previous programmes for birds (Sullivan et al., 2009), butterflies (Butterfly Conservation, 2020) and bumblebees (Bumblebee Conservation Trust, 2019) and use their established frameworks. Both citizen scientists and institutional recorders should be encouraged to record observations in secluded areas and to include observations of 'less prestigious' species.

The quality of data from, respectively, institutional recorders and citizen scientists will vary immensely depending on methods and organism group. Whereas trained professionals likely exhibit greater skills regarding some of the more challenging groups, this is not necessarily the case for all taxa. If quality can be ensured, citizen scientists can provide otherwise impossible amounts of data to facilitate science-policy impact of the sustainable biodiversity management. This study has shown the different biases from different datasets and illustrates some of the challenges with accounting for all of them in a single study.

AUTHORS' CONTRIBUTIONS

TKP, GA, JDMS and VG conceived the idea and designed the methodology; TKP retrieved and analysed the data; TKP wrote the first draft of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

All relevant data are available from public repository (GBIF Occurrence Download – 19 November 2019, doi: 10.15468/dl.dmdxne) (GBIF.org, 2019b).

Land-cover data are available through Kartkatalogen (Geonorge, 2019) and was downloaded on 23 November 2019.

All R code written to perform the data download and analyses can be viewed and downloaded in a public repository: https://doi.org/10.5281/zenodo.4455460 (Petersen, 2021).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/2688-8319.12048.

ORCID

Tanja K. Petersen https://orcid.org/0000-0002-7599-712X
James D. M. Speed https://orcid.org/0000-0002-0633-5595
Vidar Grøtan https://orcid.org/0000-0003-1222-0724
Gunnar Austrheim https://orcid.org/0000-0002-3909-6666

REFERENCES

Amano, T., Lamming, J. D. L., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *BioScience*, *66*(5), 393–400. https://doi.org/10.1093/biosci/biw022

Åsen, P. (2019) Vascular plants, field notes, Agder naturmuseum (KMN). Version 1.160. Agder Museum of Natural History and Botanical Garden. Occurrence dataset. https://doi.org/10.15468/gja4jo.

Austrheim, G., Bråthen, K. A., Ims, R. A., Mysterud, A., & Ødegaard, F. (2010). Alpine environment. In J. A. Kålås, S. Henriksen, S. Skjelseth, & Å. E. Viken (Eds.), *Environmental conditions and impacts for Red List species* (pp. 107–117). Norwegian Biodiversity Information Centre.

Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A. H., & Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS One*, *14*(9), 1–26. https://doi.org/10.1371/journal.pone.0215794

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bland, L. M., Collen, B., Orme, C. D. L., & Bielby, J. (2015). Predicting the conservation status of data-deficient species. *Conservation Biology*, *29*(1), 250–259. https://doi.org/10.1111/cobi.12372

Blindheim, T. (2020) BioFokus. version 1.1384. occurrence dataset. https://doi.org/10.15468/jxbhqx.

Boakes, E. H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D. B., & Haklay, M. (2016). Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports*, *6*, 1–11. https://doi.org/10.1038/srep33051

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bumblebee Conservation Trust (2019). *Bumblebee Conservation Trust*. https://www.bumblebeeconservation.org/.

Butterfly Conservation (2020). *Butterfly conservation*. https://butterfly-conservation.org/.

Bystriakova, N., Peregrym, M., Erkens, R. H. J., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*, 305–315. https://doi.org/10.1080/14772000.2012.705357

Chamberlain, S., & Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Chavan V., & Penev L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(S15), 1–12. https://doi.org/10.1186/1471-2105-12-s15-s2.

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, *16*(3), 354–362. https://doi.org/10.1111/j.1472-4642.2009.00615.x

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen Science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484. https://doi.org/10.1111/geb.12216

ESRI (2018) ArcGIS Desktop. Redlands: Environmental Systems Research Institute.

Gaertner, M., Wilson, J. R. U., Cadotte, M. W., MacIvor, J. S., Zenni, R. D., & Richardson, D. M. (2017). Non-native species in urban environments: Patterns, processes, impacts and challenges. *Biological Invasions*, *19*(12), 3461–3469. https://doi.org/10.1007/s10530-017-1598-7

Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodiversity Informatics*, *8*(2), 94–172. https://doi.org/10.17161/bi.v8i2.4124

Ganzevoort, W., van den Born, R. J. G., Halffman, W., & Turnhout, S. (2017). Sharing biodiversity data: Citizen scientists' concerns and motivations. *Biodiversity and Conservation*, *26*(12), 2821–2837. https://doi.org/10.1007/s10531-017-1391-z

GBIF.org (2019a). GBIF home page. https://www.gbif.org/.

GBIF.org (2019b) GBIF occurrence. Downloaded on 19 November 2019. Accessed from R via rgbif'. https://doi.org/10.15468/dl.dmdxne

Gederaas, L., Moen, T. L., Skjelseth, S., & Larsen, L.-K. (2012). *Alien species in Norway—with the Norwegian Black List 2012*. Norwegian Biodiversity Information Centre.

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, *22*(11), 1139–1149. https://doi.org/10.1111/ddi.12477

Geonorge (2019). *Kartkatalogen*. https://kartkatalog.geonorge.no/metadata/4bc2d1e0-f693-4bf2-820d-c11830d849a3.

Gjerde, I., Brandrud, T. E., Ohlson, M., & Ødegaard, F. (2010). Woodland. In J. A. Kålås, S. Henriksen, S. Skjelseth, & Å. E. Viken (Eds.), (Eds.) *Environmental conditions and impacts for Red List species* (pp. 67–78). Norwegian Biodiversity Information Centre.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*, 1424–1435. https://doi.org/10.1111/ele.12189

Henriksen, S., & Hilmo, O. (2015) Norwegian Red List of Species 2015 methods and results. Norwegian Biodiversity Information Centre, Norway ISBN: 978-82-92838-44-0

Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., & Ono, S. (2015). Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions*, *21*(1), 46–54. https://doi.org/10.1111/ddi.12255

Hsieh, T. C., Ma, K. H., & Chao, A. (2020) iNEXT: INterpolation and EXTrapolation for species diversity. R package version 2.0.20.

IUCN (2020). *Invasive species*. https://www.iucn.org/theme/species/our-work/invasive-species.

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology and Evolution*, *27*(3), 151–159. https://doi.org/10.1016/j.tree.2011.09.007

Jordal, J. B. (2019) *Jordal. version 1.91. Biolog J.B. Jordal AS. Occurrence dataset*. https://doi.org/10.15468/wqsad9.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364

Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Larson, L. R., Cooper, C. B., Futch, S., Singh, D., Shipley, N. J., Dale, K., LeBaron, G. S., & Takekawa, J. Y. (2020). The diverse motivations of citizen scientists: Does conservation emphasis grow as volunteer participation progresses?. *Biological Conservation*, *242*, 108428. https://doi.org/10.1016/j.biocon.2020.108428

Lehikoinen, A., Brotons, L., Calladine, J., Campedelli, T., Escandell, V., Flousek, J., Grueneberg, C., Haas, F., Harris, S., Herrando, S., Husby, M., Jiguet, F., Kålås, J. A., Lindström, Å., Lorrillière, R., Molina, B., Pladevall, C., Calvi, G., Sattler, T., … Trautmann, S. (2019). Declining population trends of European mountain birds. *Global Change Biology*, *25*(2), 577–588. https://doi.org/10.1111/gcb.14522

Levatich, T., & Padilla, F. (2019) *EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset*. https://doi.org/10.15468/aomfnb.

Luck, G. W. (2007). A review of the relationships between human population density and biodiversity. *Biological Reviews*, *82*(4), 607–645. https://doi.org/10.1111/j.1469-185X.2007.00028.x

Magurran, A. E., Dornelas, M., Moyes, F., & Henderson, P. A. (2019). Temporal *β* diversity—A macroecological perspective. *Global Ecology and Biogeography*, *28*(12), 1949–1960. https://doi.org/10.1111/geb.13026

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285–290. https://doi.org/10.1890/110278

Natural History Museum, U. of O. (2019a) *Vascular Plants, Field notes, Oslo (O). Version 1.186. Occurrence dataset*. https://doi.org/10.15468/w8gru5.

Natural History Museum, U. of O. (2019b) *Vascular Plants, Observations, Oslo (O). Version 1.181. Occurrence dataset*. https://doi.org/10.15468/tvnjk7.

Natural History Museum, U. of O. (2020) *Lichen field notes, Oslo (O). Version 1.180. Occurrence dataset*. https://doi.org/10.15468/zrfxcu.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, *34*(1), 3–22. https://doi.org/10.1177/0309133309355630

The Norwegian Biodiversity Information Centre (2018a). *Kvalitetssikring og foredling av dataene*. https://artsdatabanken.no/Pages/233743/Kvalitetssikring_og_foredling_av_dataene.

Norwegian Institute of Bioeconomy Research (2019). AR50. https://nibio.no/tema/jord/arealressurser/ar50.

Norwegian Species Observation Service (2020). *Håndtering av avvikende rapporter i Artsobservasjoner*. https://www.artsobservasjoner.no/Home/DeviatingReports.

NTNU University Museum (2018) Samlingsplan 2018 - 2025, Samlingsplan NTNU Vitenskapsmuseet. Trondheim, Norway.

NTNU University Museum (2020). *Vascular plant field notes, NTNU University Museum. Version 1.97. Sampling event dataset*. https://doi.org/10.15468/kkb2x0.

Ogle, D. H., Wheeler, P., & Dinno, A. (2020). Fisheries Stock Analysis. R package version 0.8.30.

Osawa, T. (2019). Perspectives on biodiversity informatics for ecology. *Ecological Research*, *34*(4), 446–456. https://doi.org/10.1111/1440-1703.12023

Pärtel, M., Bruun, H. H., & Sammul, M. (2005). Biodiversity in temperate European grasslands: Origin and conservation. *Grassland Science in Europe*, *10*, 1–14.

Pebesma, E., & Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13.

Petersen, T. K. (2021). tanjakofodpetersen/Species-data-for-understanding-biodiversity-dynamics: Code for analyses (Version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4455460

Phillips, S. J., Dudík, M., Dudík, D., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.

Pohlert, T. (2020) trend: Non-Parametric Trend Tests and Change-Point Detection. R package version 1.1.2.

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019). Ecological data should not be so hard to find and reuse. *Trends in Ecology and Evolution*, *34*(6), 494–496. https://doi.org/10.1016/j.tree.2019.04.005

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*(3), 532–542. https://doi.org/10.1111/bij.12517

Core Team, R. (2020). *R: A language and environment for statistical computing'*. R Foundation for Statistical Computing.

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*(4), 460–472. https://doi.org/10.1111/ddi.12698

Salomon Cavin, J., & Kull, C. A. (2017). Invasion ecology goes to town: From disdain to sympathy. *Biological Invasions*, *19*(12), 3471–3487. https://doi.org/10.1007/s10530-017-1588-9

Sandvik, H., Gederaas, L., & Hilmo, O. (2017). Guidelines for the generic ecological impact assessment of alien species. *version 3*. Norwegian Biodiversity Information Centre.

Speed, J. D. M., Bendiksby, M., Finstad, A. G., Hassel, K., Kolstad, L., & Prestø, T. (2018). Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLoS Biology*, *13*(4), 1–17. https://doi.org/10.1371/journal.pone.0196417

Statistics Norway (2020). *Statistisk Sentralbyrå*. https://www.ssb.no/.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Syfert, M. M., Joppa, L., Smith, M., Coomes, D., Bachman, S., & Brummitt, N. (2014). Using species distribution models to inform IUCN Red List assessments. *Biological Conservation*, *177*, 174–184.

The Norwegian Biodiversity Information Centre (2015). *Norwegian Red List for species*. https://www.biodiversity.no/Pages/135380/Norwegian_Red_List_for_Species?Key=14.

The Norwegian Biodiversity Information Centre (2018b). *The alien species list of Norway*. https://www.biodiversity.no/alien-species.

The Norwegian Biodiversity Information Centre and Hoem, S. (2020a) *Norwegian Biodiversity Information Centre - Other datasets. Version 13.132. Occurrence dataset*. https://doi.org/10.15468/tm56sc.

The Norwegian Biodiversity Information Centre and Hoem, S. (2020b) *Norwegian Species Observation Service. Version 1.82. Occurrence dataset*. The Norwegian Biodiversity Information Centre (NBIC). https://doi.org/10.15468/zjbzel.

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A., & Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Thuiller, W., Richardson, D. M., Pyšek, P., Midgley, G. F., Hughes, G. O., & Rouget, M. (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, *11*, 2234–2250. https://doi.org/10.1111/j.1365-2486.2005.01018.x

Tiago, P., Gouveia, M. J., Capinha, C., Santos-Reis, M., & Pereira, H. M. (2017). The influence of motivational factors on the frequency of participation in citizen science activities. *Nature Conservation*, *18*, 61–78. https://doi.org/10.3897/natureconservation.18.13429

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-09084-6

Tulloch, A. I. T., Mustin, K., Possingham, H. P., Szabo, J. K., & Wilson, K. A. (2013). To boldly go where no volunteer has gone before: Predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, *19*(4), 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x

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. https://doi.org/10.1016/j.biocon.2013.05.025

Tye, C. A., McCleery, R. A., Fletcher, R. J., Greene, D. U., & Butryn, R. S. (2017). Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, *54*(2), 628–637. https://doi.org/10.1111/1365-2664.12682

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

Yañez-Arenas, C., Guevara, R., Martínez-Meyer, E., Mandujano, S., & Lobo, J. M. (2014). Predicting species' abundances from occurrence data: Effects of sample size and bias. *Ecological Modelling*, *294*, 36–41. https://doi.org/10.1016/j.ecolmodel.2014.09.014

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Petersen TK, Speed JDM, Grøtan V, Austrheim G. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecol Solut Evidence*. 2021;2:e12048. https://doi.org/10.1002/2688-8319.12048