ECOGRAPHY

Research

Temporal trends in the spatial bias of species occurrence records

Diana E. Bowler, Corey T. Callaghan, Netra Bhandari, Klaus Henle, M. Benjamin Barth, Christian Koppitz, Reinhard Klenke, Marten Winter, Florian Jansen, Helge Bruelheide and Aletta Bonn

D. E. Bowler (https://orcid.org/0000-0002-7775-1668) ✉ (diana.e.bowler@gmail.com), C. T. Callaghan (https://orcid.org/0000-0003-0415-2709), N. Bhandari, R. Klenke, M. Winter (https://orcid.org/0000-0002-9593-7300), H. Bruelheide (https://orcid.org/0000-0003-3135-0356) and A. Bonn (https://orcid.org/0000-0002-8345-4600), German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. DEB and AB also at: Friedrich Schiller Univ. Jena, Inst. of Biodiversity, Jena, Germany. – DEB, AB and K. Henle, Helmholtz-Center for Environmental Research – UFZ, Dept Ecosystem Services, Leipzig, Germany. CTC, RK and HB also at: Martin Luther Univ. Halle-Wittenberg, Inst. of Biology/Geobotany and Botanical Garden, Halle, Germany. – M. Benjamin Barth, Leipzig Rural District Office, Environmental Agency, Section of Nature and Landscape Conservation, Borna, Germany. – C. Koppitz, Landesamt für Landwirtschaft, Umwelt und Ländliche Räume des Landes Schleswig-Holstein, Flintbek, Germany. – F. Jansen (https://orcid.org/0000-0002-0331-5185), Univ. of Rostock, Faculty of Agricultural and Environmental Sciences, Rostock, Germany.

Ecography
2022: e06219
doi: 10.1111/ecog.06219
Subject Editor: Pedro Peres-Neto
Editor-in-Chief: Miguel Araújo
Accepted 6 April 2022

Large-scale biodiversity databases have great potential for quantifying long-term trends of species, but they also bring many methodological challenges. Spatial bias of species occurrence records is well recognized. Yet, the dynamic nature of this spatial bias – how spatial bias has changed over time – has been largely overlooked. We examined the spatial bias of species occurrence records within multiple biodiversity databases in Germany and tested whether spatial bias in relation to land cover or land use (urban and protected areas) has changed over time. We focused our analyses on urban and protected areas as these represent two well-known correlates of sampling bias in biodiversity datasets. We found that the proportion of annual records from urban areas has increased over time while the proportion of annual records within protected areas has not consistently changed. Using simulations, we examined the implications of this changing sampling bias for estimation of long-term trends of species' distributions. When assessing biodiversity change, our findings suggest that the effects of spatial bias depend on how it affects sampling of the underlying land-use change drivers affecting species. Oversampling of regions undergoing the greatest degree of change, for instance near human settlements, might lead to overestimation of the trends of specialist species. For robust estimation of the long-term trends in species' distributions, analyses using species occurrence records may need to consider not only spatial bias, but also changes in the spatial bias through time.

Keywords: biodiversity change, biodiversity monitoring, citizen science, opportunistic data, presence-only data

Introduction

Quantifying population and distribution trends in space and time is increasingly important for biodiversity monitoring, conservation and related political decisions (Yoccoz et al. 2001, Harrison et al. 2014). Species occurrence records have great

NORDIC SOCIETY OIKOS
www.ecography.org

potential for biodiversity monitoring because of their large taxonomic, geographic and temporal scope (Chandler et al. 2017, Fink et al. 2020). Occurrence record databases comprise different types of data, ranging from opportunistic (i.e. incidental observations) to standardized (e.g. collected following protocols and sampling design) (Isaac and Pocock 2015); however, often all the data are treated as a single data type ('presence-only'). Species occurrence records have been used for many studies on the spatial patterns of species' distributions (Feeley and Silman 2011, Beck et al. 2014, Sullivan et al. 2017, Bradter et al. 2018), and are increasingly used in studies about temporal trends (Powney et al. 2019, Outhwaite et al. 2020, Bowler et al. 2021, Sheard et al. 2021, Zattara and Aizen 2021). However, the data are often spatially biased (Geldmann et al. 2016), raising concerns about their validity and use in ecological research (Burgess et al. 2017, Bayraktarov et al. 2019). Quantifying the extent of these spatial biases, and especially whether the biases have changed through time, is important for the continued use of occurrence records in studies about temporal change of biodiversity.

Spatial bias in species occurrence records can be seen at a range of different spatial scales (Boakes et al. 2010, Amano et al. 2016, Freeman and Peterson 2019, Girardello et al. 2019). At the global-scale, more occurrence records are found in countries with developed biodiversity data infrastructure and higher GDP (Collen et al. 2008, Meyer et al. 2015, Callaghan et al. 2021c). At regional or national-scales, spatial bias is commonly associated with human population density, settlements and roads (Kelling et al. 2015, Geldmann et al. 2016, Mair and Ruete 2016, Hugo and Altwegg 2017, Dissanayake et al. 2019, Girardello et al. 2019, Husby et al. 2021). The higher number of biodiversity records near urban areas can reflect the local densities of observers as well as the behavior of observers to visit the most accessible places from their home. Many occurrence records are also collected in protected areas or hotspots of species richness (Geldmann et al. 2016, Hugo and Altwegg 2017, Jimenez-Valverde et al. 2019). All of these biases potentially limit the use of the data within heterogeneous occurrence record databases compared with data from a more structured stratified sampling design (Buckland and Johnston 2017).

While these spatial biases, and their influence on ecological models, are widely recognized (Beck et al. 2014, Hugo and Altwegg 2017, Fournier et al. 2019), most studies have focused on temporally static patterns of bias. Hence, few studies have examined whether, and to what extent, spatial biases have changed over time. Annual numbers of occurrence records within databases such as the Global Biodiversity Information Facility (GBIF), however, have massively increased over time (Boakes et al. 2010, Petersen et al. 2021). Additionally, 70–80% of all species occurrence data in Europe are estimated to have been collected by volunteers (Schmeller et al. 2009). The increasing number of occurrence records reflects a combination of increased awareness and participation of citizen science (defined here as any voluntary data collection – recording species either opportunistically or by following a standardized protocol) and new technologies for recording and submitting species observations (Chandler et al. 2017, Mihoub et al. 2017), as well as mobilization of other data sources, such as museum and literature records (Boakes et al. 2010). Such increasing numbers of records might also be associated with changes in the spatial bias in the data. In a butterfly dataset in the United States, for example, inventory completeness was greater in regions of high human density and this association became stronger over time (Shirey et al. 2021). Changes in the strength and pattern of spatial biases of species occurrence records could occur for many reasons – land use change at sites already being monitored, changes in where people collect data, and/or changes in the types or behaviours of people and projects collecting and reporting data (Isaac and Pocock 2015, August et al. 2020, Petersen et al. 2021). As species occurrence data are being increasingly used to estimate species' long-term trends, there is a danger of inferences being affected by changing spatial bias.

We studied the spatial bias of species occurrence data and the implications of changing spatial biases for biodiversity change research, using a combination of empirical analysis of species occurrence record databases and simulations. We focused on spatial bias towards urban and protected areas since previous studies have documented spatial bias with respect to these land covers/uses (Kelling et al. 2015, Geldmann et al. 2016, Mair and Ruete 2016, Hugo and Altwegg 2017, Dissanayake et al. 2019, Girardello et al. 2019, Husby et al. 2021). We used multiple datasets from different databases, which include citizen science and other sources of species occurrence records, for birds, amphibians, butterflies and plants. For each dataset, we quantified spatial sampling bias in relation to each type of land cover/use and tested whether the strength of the bias had changed through time. Using simulations, we then explored the effects of changing spatial bias on estimated species' distribution trends, assuming different sampling scenarios, patterns of environmental change and species' habitat associations.

Methods

Empirical study

Biodiversity data

We used three datasets differing in the proportion of records from citizen science. These datasets came from 1) GBIF (<www.gbif.org/>); 2) Naturgucker (<www.naturgucker.de>) and 3) Observation.org (<https://observation.org/>). The latter two platforms are specifically targeted towards citizen scientists while GBIF includes a mix of records from citizen science (also from these platforms), as well as data from museums and research institutions. In each dataset, we retrieved species occurrence records for selected taxa: amphibian/reptiles; birds; butterflies/moths and vascular plants. We

focused on these groups because they are popular targets of citizen science. For simplicity, we subset the datasets to records collected within a three-month period in the year, based on when the number of records peaked (April–June for all except butterflies/moths, which peaked June–August). Additionally, for amphibians/reptiles, we used a dataset of occurrence records collected in Saxony, an administrative region (federal state) of Germany, compiled from the Central Species Database of Saxony by the regional conservation agency (<www.natur.sachsen.de/zentrale-artdatenbank-zena-sachsen-6905.html>), which contained both opportunistic and systematic monitoring data. For each dataset, species occurrence data were mapped to quadrants of the German ordnance survey (TK25 quadrant, ca 5.5 × 6 km) that is commonly used in Germany to define the location of species observations. Hence, we defined a site as a TK25 quadrant. For all datasets, we used available species occurrence data between 1992 and 2018 to align with the available land cover data.

Land-use data

We used the European Space Agency Climate Change Initiative land cover dataset (ESA 2017), providing annual data at a 300 m resolution for each year between 1992 and 2018. We calculated for each year and TK25 quadrant the proportional cover of urban cover. For spatial information on protected areas, we used GIS shapefiles from the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz, BfN) – focusing on the protected area categories of nature reserves (Naturschutzgebiet) and national parks (Nationalparks) that have the highest level of protection in Germany. In 2018, there were 8833 nature reserves and 16 national parks across Germany.

Statistical analysis

To examine spatial bias, we tested the effect of land cover/use on the probability of a site being visited in each dataset. In the datasets, people report species observations and not specifically when a site has been visited for a survey. We defined a site as visited when there was at least one species occurrence record in a site within a year, which indicated that some kind of survey by a person had taken place. While there may also be within-year seasonal variation in visitation patterns in relation to land cover/use, we decided to focus only on annual patterns since that was sufficient for our purpose and most relevant for questions about long-term trends focusing on between-year changes. We first used a generalized linear model (GLM) with binomial error distribution to analyze the main effect of urban cover and protected area on whether a site was visited (a binary variable, yes or no) across all years. Under random sampling, site visit probability should be independent of land cover/use (i.e. a regression coefficient close to zero). We then examined the evidence for changing spatial bias by including year and the interaction between land cover/use (urban and protected area) and year on whether a site was visited. Under the null assumption of no change in spatial bias through time, there should be no significant interactions between land cover/use and year on the probability of site visitation.

Changes in spatial bias with respect to land cover/use could be caused by multiple underlying processes, including 1) expansion of surveys to new sites, e.g. by new recorders; 2) no expansion but rather a shift in the frequency with which different sites are sampled; or 3) faster changes in land cover/use at sites that are already being sampled (see Supporting information for details). To examine the evidence for these different possibilities, we calculated the Pearson correlation coefficients between the estimate of spatial bias in a given year and the total number of recorders and sites surveyed, as well as the proportion of new sites or recorders (according to their first year of survey). Recorder information was not available for the Observation.org dataset. We also explored the relationship between urbanization (urban cover change during 1992–2018) and the proportion of years (two-column response) in which a site was sampled using a binomial GLM.

We repeated all the models while accounting for spatial autocorrelation, based on a Matérn correlation structure using the geographic coordinates of each TK25 quadrant, using the spaMM R package (Rousset and Ferdy 2014). We used a Matérn correlation structure because our coordinates were arranged in a regular grid and this structure allowed the spatial autocorrelation to depend on the distance between them. Ideally, for most ecological questions, spatial bias should be not present regardless of the spatial pattern, but spatial autocorrelation of visits may contribute to the estimated spatial bias. In fact, similar patterns were found. In the main text, we focus on the results for the largest datasets (GBIF, Naturgucker) but show the results for all in the SI based on the models that accounted for spatial autocorrelation.

Simulations

While our empirical study informed on the evidence for changing spatial bias, we used simulations to examine the broader implications of changing spatial bias for studies using these types of data sources to infer long-term species population changes. We compared the type of sampling bias observed in our empirical analysis with alternative sampling scenarios. We also tested the effects under different scenarios for land-use change and species habitat preference.

Species dynamics

We assumed a landscape of 500 sites that varied in urban cover. For convenience, we assumed urban cover to be uniformly distributed among sites between −1 and 1, where −1 indicated low urban cover and 1 indicated high urban cover. We could have scaled urban cover between 0 and 100% but this would not change our results. We modelled the dynamics of one species whose occupancy probability was affected by urban cover and had an occupancy probability of 0.5 at mean urban cover (represented by a zero value). For our simulation, we initially assumed that a species was an urban avoider, with a negative effect (coefficient of −2 on the logit-scale) of urban cover on its occupancy, and we therefore present the

results of the urban avoider in the main results. However, in further analyses, we varied the species' habitat preference by also assuming a species was an urban exploiter and urban generalist (see Table 1 – 'Species habitat preference'). A negative assumption was selected for the main analysis since there are generally more urban avoiders than exploiters within a given taxon group (Callaghan et al. 2021a, b). Occupancy states for each site were drawn from a Bernoulli distribution of the occupancy probability to determine species presence or absence. For simplicity, urban cover was assumed to be the only deterministic factor affecting species occupancy; hence, urban cover change also caused species occupancy change.

Environmental change

We assumed two time-points during which urban cover within a grid changed according to different change scenarios (Table 1, 'Environmental (urban) change scenarios'). Urban change was either uniform (all sites increased in urban cover by the same amount) or clustered (regional urban cover increased by the same total amount, but it was concentrated in the sites with above-average urban cover). Clustered change was intended to mimic urban expansion within and at the edge of urban areas. Clustered change is more typical of real-world contexts due to spatial heterogeneity in pressures, but we included uniform change to compare the effect of different change patterns.

Sampling scenarios

We created different spatial sampling scenarios to reflect urban sampling bias, and temporal increase in urban sampling bias, and contrasted these with random sampling (Fig. 1). We assumed that each site was visited at each time point according to some probability that varied among different sampling scenarios (see Table 1, 'Sampling scenarios'). For simplicity, we assumed perfect detection, i.e. if a site was visited and the species was present, then the species was always detected. In sampling scenarios *full* and *random*, sites were sampled fully (i.e. all sites visited) or a proportion of sites was visited at random (20% of sites – to assume a moderately low level of sampling typical of real world sampling), respectively. In scenarios *bias* and *bias+*, sites were not sampled at random; instead, site visit probability increased linearly (coefficient of 2 on the logit-scale) with urban cover. In scenario *bias+*, this coefficient increased by a factor of 3 in the second time point, i.e. sampling bias increased through time. Site visit (yes or no) was then determined by Bernoulli draws of the site visit probability. Across all sampling scenarios and time-points, we controlled for the total number of visited sites (20% of sites except for *full*) and kept site visit probability at 50% at the mean urban cover within a landscape to control for total sampling effort. We ran 1000 replications for each set of scenario combinations. *Bias+* corresponded to the pattern observed in our empirical datasets.

Statistical analysis

We calculated the occupancy proportion (number of sampled sites in which the species was present) for each time point and simulation replicate. We also calculated the occupancy change as the differences in the log-odds of occupancy between time points. We used these sample estimates as the best estimates for species occupancy patterns and changes. We did not attempt to control for the sampling bias since we were interested in understanding the implications of a naive analysis. We also ran formal models for hypothesis tests of the evidence for species occupancy change, using generalized linear mixed effects models (GLMM) with binomial error distribution, including species observation (present or absent) as the response, year as the explanatory variable and site as a random effect. Using these models, we calculated the Type I error (number of significant ($p < 0.05$) year effects under no urban or species occupancy change). We used R ver. 4.0.2 (<www.r-project.org>) for simulations and analysis.

Results

Empirical patterns

In both GBIF and Naturgucker datasets, sites with higher urban cover were, on average, more likely to be visited than

Table 1. Simulation scenarios.

Sampling scenarios:	
Full	All sites sampled
Random	20% sampled – at random
Bias	20% sampled – urban sites overrepresented (constant spatial bias)
Bias+	20% sampled – urban sites overrepresented (increasing spatial bias)
Environmental (urban) change scenarios:	
No change	No urban cover change
Uniform change	Uniform increase in urban cover (all grids increase in urban cover)
Clustered change	Clustered increase in urban cover (only grids with above-average urban cover increase in urban cover)
Species' habitat preference	
Urban exploiter	Species occurrence probability is positively associated with urban cover
Generalist	Species occurrence probability is not associated with urban cover
Urban avoider	Species occurrence probability is negatively associated with urban cover (main assumption)

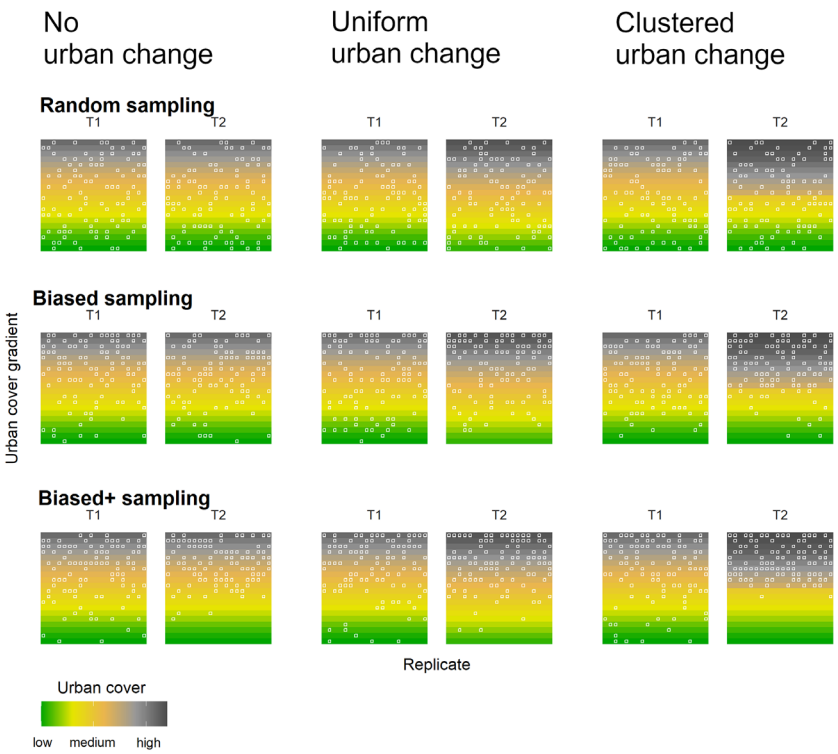

Figure 1. Schematic figure of the simulation scenarios in theoretical landscapes assuming two time points T1 and T2 with varying degrees of urban cover change between T1 and T2 (no urban cover change, uniform urban cover change or clustered urban change). Sample locations in each simulation replicate are shown by the white squares. Sites are either sampled at random, or spatially biased (*bias*) or increasingly spatially biased through time (*bias+*) towards urban cover. The colour scale from grey through orange to green indicates a gradient from high to low urban cover.

sites with lower urban cover, and this pattern became stronger over time (Fig. 2, Supporting information). Overall, mean site visit probability increased between 1992 and 2018, consistent with the increase in the number of submitted occurrence records (Supporting information). However, visit probability increased more strongly for sites with high urban cover than for sites with low urban cover, leading to significant interactions between urban cover and year in most cases (Fig. 2, Supporting information). Similar patterns were found in the regional conservation agency data for amphibians (Supporting information) but were weaker in the Observation.org data, which contained the smallest number of records (Supporting information).

Across all years, visit probability also tended to be spatially biased towards protected areas within both the GBIF and Naturgucker datasets (Fig. 3, Supporting information). However, the strength of the bias towards protected areas did not consistently change through time (Fig. 3, Supporting information). Again, similar patterns were found in the regional conservation agency data for amphibians (Supporting information) and the Observation.org data (Supporting information).

Different processes might explain the increasing spatial bias towards urban areas (Supporting information). Annual bias estimates towards urban cover were most strongly associated with the total numbers of recorders ($r > 0.5$ in 8/8 datasets) and total number of sites surveyed ($r > 0.5$ in 10/12 datasets). But we found only weak associations between annual bias and the proportion of new recorders or new sites in a given year (Supporting information). There was also a positive effect of urban cover change on the proportion of years in which a site was sampled, i.e. sites undergoing urbanization were visited in a higher proportion of the years between 1992 and 2018 (Supporting information).

Simulations

We first present detailed results for a theoretical species that is negatively associated with urban land cover, and then later show effects for all three types of species that are negatively, neutral or positively associated with urban land cover.

Changes under no urban change

Under the 'no urban change' scenario, true species occupancy did not change between the time points. Sampling all sites (*full*) or random sampling (*random*) led to the same mean proportion of occupied sites (Fig. 4A) and no change in the proportion of occupied sites between time points (Fig. 4B). With sampling bias (*bias* and *bias+*), the estimated proportion of occupied sites was underestimated because the species was more common at low urban cover sites while sites with high urban cover were preferentially sampled (Fig. 4A). However, provided the sampling bias did not change over

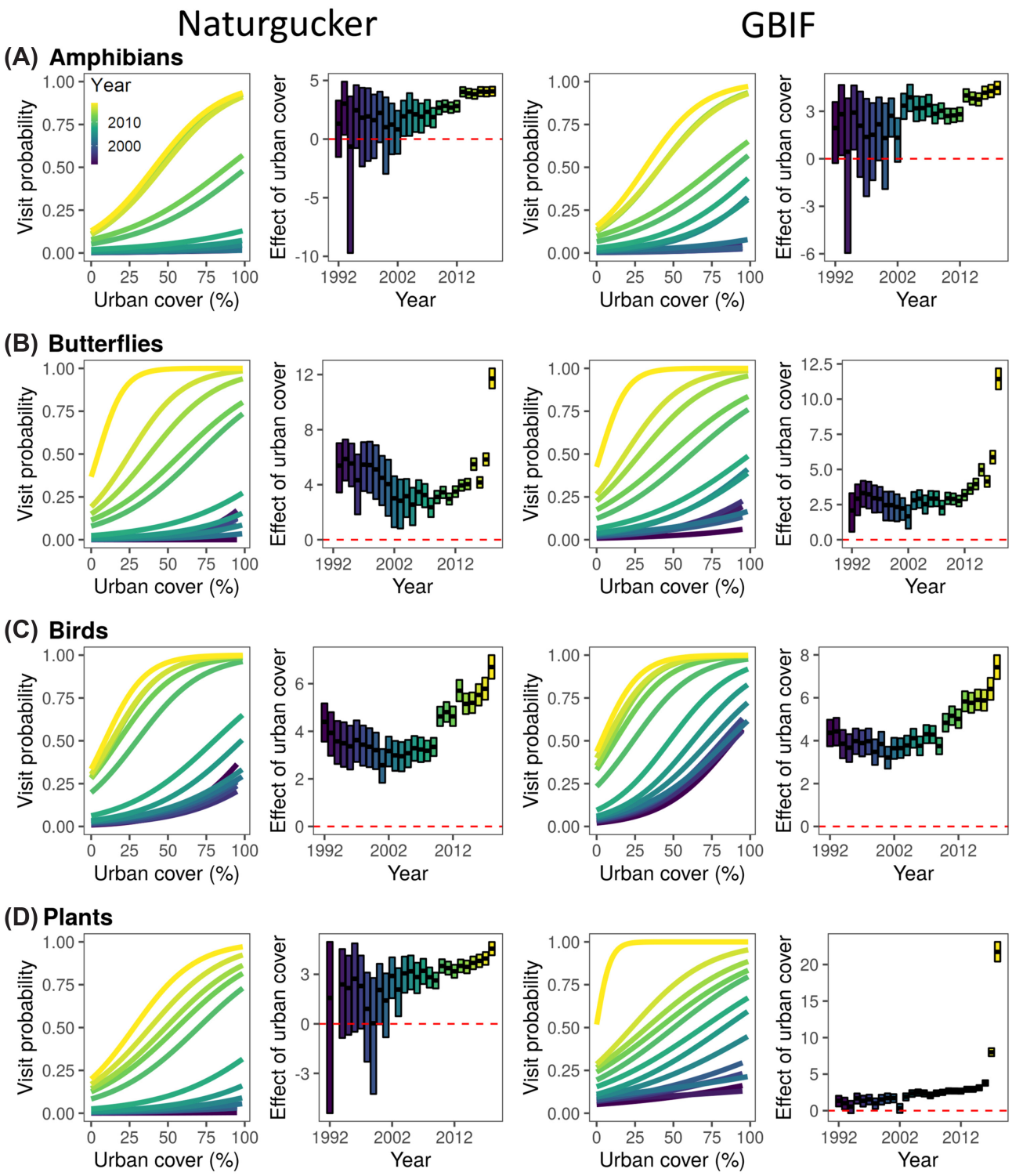

Figure 2. The effect of urban cover on the probability that a site is visited within a given year in each empirical dataset. Lines are coloured to reflect the order of the years from older (purple = 1992) to more recent (yellow = 2018). Left: each line shows the relationship between urban cover and site visit probability (grouped into two-years to simplify visualization) as predicted by a binomial GLM. Right: the effect of urban cover (logit-scale) on site visit probability as estimated by a binomial GLM in each year. Bars show the mean and 95% confidence intervals of the estimated effects. The dashed line is the line of no effect.

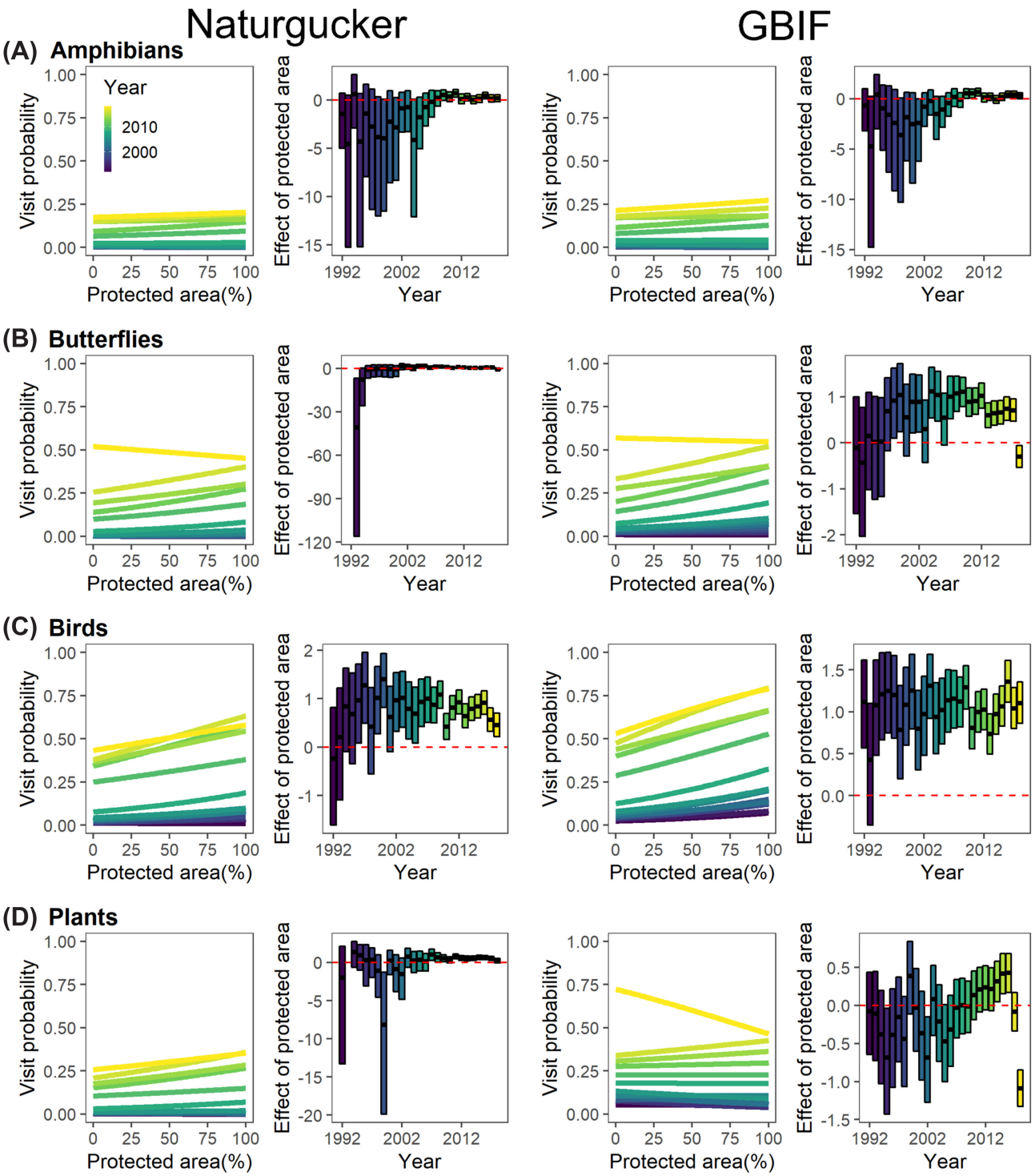

Figure 3. The effect of protected area on the probability that a site is visited within a given year in each empirical dataset. Lines are coloured to reflect the order of the years from older (purple = 1992) to more recent (yellow = 2018). Left: each line shows the relationship between protected area and site visit probability (grouped into two-years to simplify visualization) as predicted by a binomial GLM. Right: the effect of protected area (logit-scale) on site visit probability as estimated by a binomial GLM in each year. Bars show the mean and 95% confidence intervals of the estimated effects. The dashed line is the line of no effect.

time (scenario *bias*), sampling still yielded the correct assessment of no change in the proportion of occupied sites between the time points (Fig. 4B). By contrast, when sampling bias increased between the time points (scenario *bias*+), inferred change in the proportion of occupied sites was biased and the species was estimated to decline (Fig. 4B). Testing the significance of the occupancy change resulted in greater Type I error (i.e. false positives over 20% of simulations) under *bias*+ (Supporting information) compared to the other sampling scenarios (less than 5% false positives).

Changes under urban change

Urbanization between time points caused a decline in the occurrence of the species, as assumed by the underlying model of an urban avoiding species (Fig. 4B). The effect of constant or increasing sampling bias on the estimated occupancy change depended on whether urban change was uniform across all sites (i.e. if all sites increased in urban cover) or clustered (i.e. if only sites with medium or high urban cover increased in urban cover). Constant spatial bias (*bias*) under a uniform pattern of urban change led to similar estimates of change as random sampling, with species occupancy declining. However, constant spatial bias led to an overestimation of the decline under a clustered pattern of urban change (Fig. 4). Increasing spatial bias (*bias*+) led to an overestimation of species declines under both uniform and clustered patterns of urban change (Fig. 4).

Overall, the patterns can be explained by how each sampling scenario captured the distribution of urban change in the landscape, which was driving species occupancy change (Fig. 5). Under no or uniform urban change, sampling randomly or with a constant spatial bias both captured urban change in a representative way. Mean sampled urban cover was higher under constant biased sampling than random sampling; however, the change in sampled urban cover between the time points was similar for both (Fig. 5). This explains why both sampling scenarios led to similar estimates of species occupancy change (Fig. 4B). With increasing spatial sampling bias, urban cover at the sampling sites increased at the second time point, leading to overly high estimates of species declines (Fig. 4B).

Under clustered urban change, both constant and increasing spatial sampling bias resulted in unrepresentative samples of urban change – shown by mean urban cover at sampling sites increasing between time points to a greater extent than with random sampling (Fig. 5). Hence, under both these sampling scenarios, urban change and species occupancy change were oversampled, explaining the overestimated species declines (Fig. 4B).

Effect of species' habitat preferences

In the previous results, we assumed that species occupancy declined as urban cover increased between the two time points ('urban avoider'). We also ran additional simulations

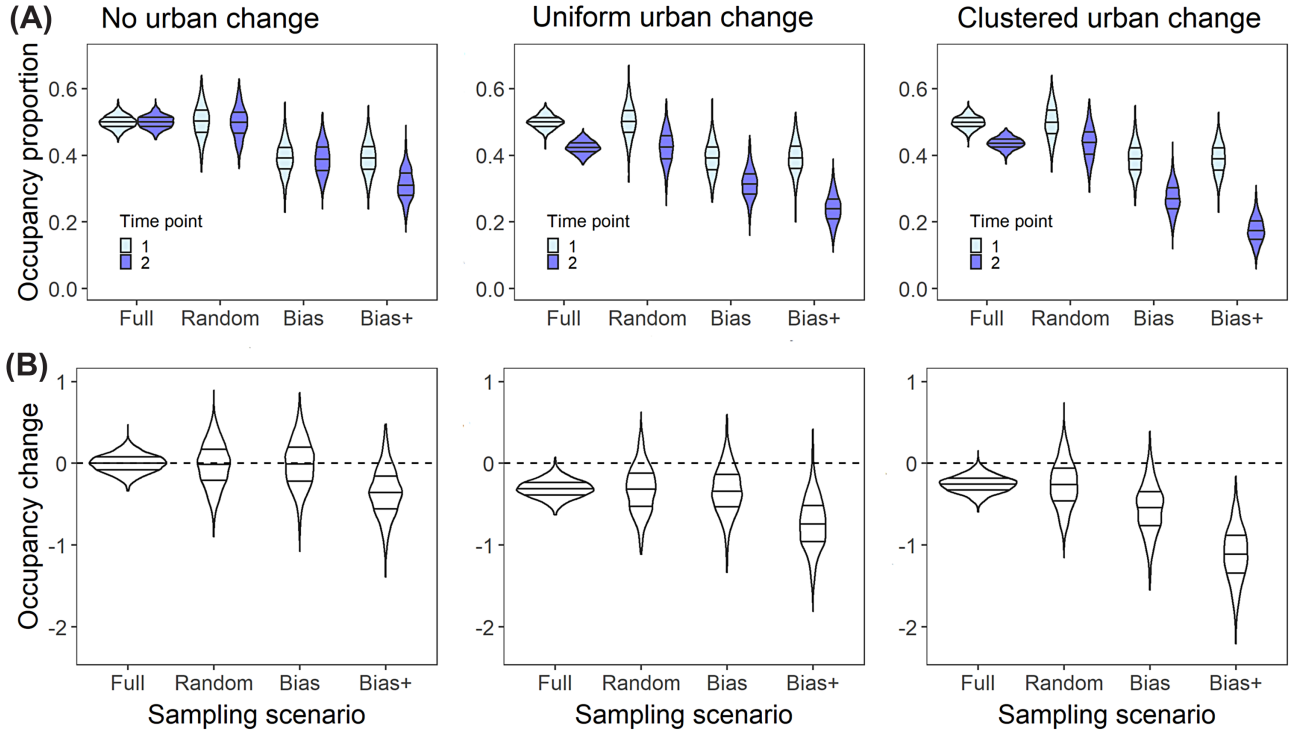

Figure 4. Violin plots showing the effect of each sampling scenario and urban change pattern on (A) the observed proportion of occupied sites and (B) the observed change in occupancy (log-odds ratio of occupancy proportion between the two time points) across 1000 simulations. In all scenarios, the species was assumed to be negatively affected by urban cover. Scenario *full* was when all sites are sampled and hence represents the truth. Scenario *random* was random sampling. Scenario *bias* was constant spatial bias (towards urban cover). Scenario *bias*+ was increasing spatial bias (towards urban cover) between the time points. The dashed horizontal line in B represents the line of no change in occupancy proportion between the two time points.

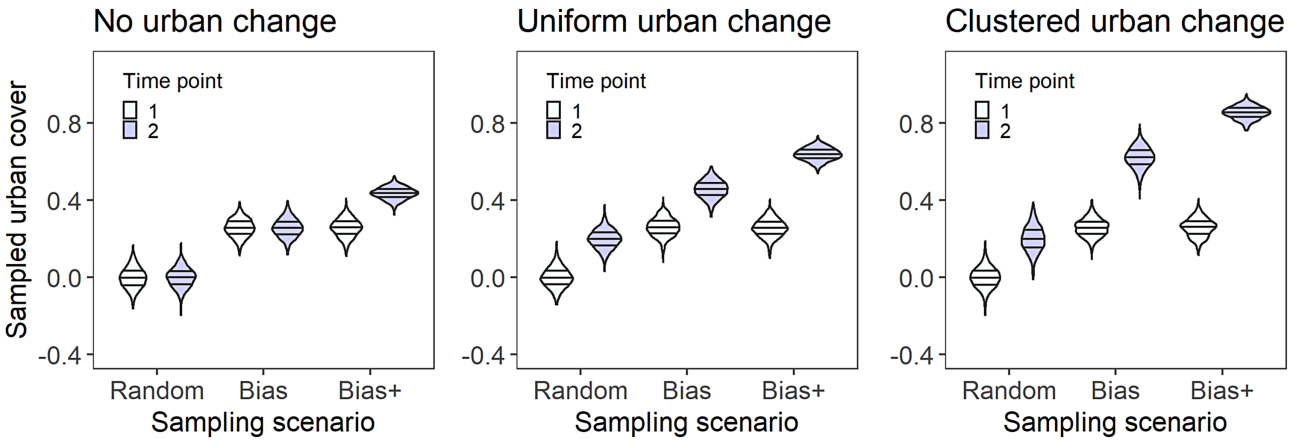

Figure 5. Violin plots showing the sampled urban cover (mean urban cover across sampling sites) at each time point for each sampling scenario and environmental change scenario across 1000 simulations. Scenario *random* was random sampling. Scenario *bias* was constant spatial bias (towards urban cover). Scenario *bias+* was increasing spatial bias (towards urban cover).

to explore the effects of different species' habitat preferences (Fig. 6). Under no or uniform urban change, species association did not matter provided that the sampling bias was constant through time (*bias*). Under all other scenarios (increasing bias, *bias+*, in all urban change patterns, or constant bias with clustered urban change), the strength of the bias in the estimated occupancy change increased with the strength of species habitat association.

For urban exploiters (species with a positive association with urban cover), this meant that species increases were overestimated; while for urban avoiders, declines were overestimated. For a generalist species (species with no association with urban cover), neither sampling bias nor the pattern of urban change had any effect on the estimated occupancy change.

Discussion

Databases of species occurrence records are increasingly used to quantify large-scale biodiversity patterns in space and time (Theobald et al. 2015, Chandler et al. 2017). Many studies have already raised concerns associated with spatial biases in these databases (Geldmann et al. 2016, Girardello et al. 2018) and begun to develop solutions to account for spatial biases (Johnston et al. 2020). Here, we highlight an overlooked form of bias – changing spatial bias through time – with implications for biodiversity change assessments using species occurrence records collected without a coordinated sampling design. Specifically, we found evidence that sampling bias towards urban areas has become stronger in recent years. While we focused on German datasets of species occurrences, similar data are available for many other countries; hence, we expect to find the same patterns elsewhere. Our simulations suggest that effects of spatial sampling bias depend on how it affects sampling of the underlying land-use drivers of species trends. Biased estimates of species population change arise when the underlying drivers of species change are not representatively sampled.

Changing spatial sampling bias

Taxonomic and geographic biases of databases are well-studied (Meyer et al. 2015, Troudet et al. 2017), but few studies

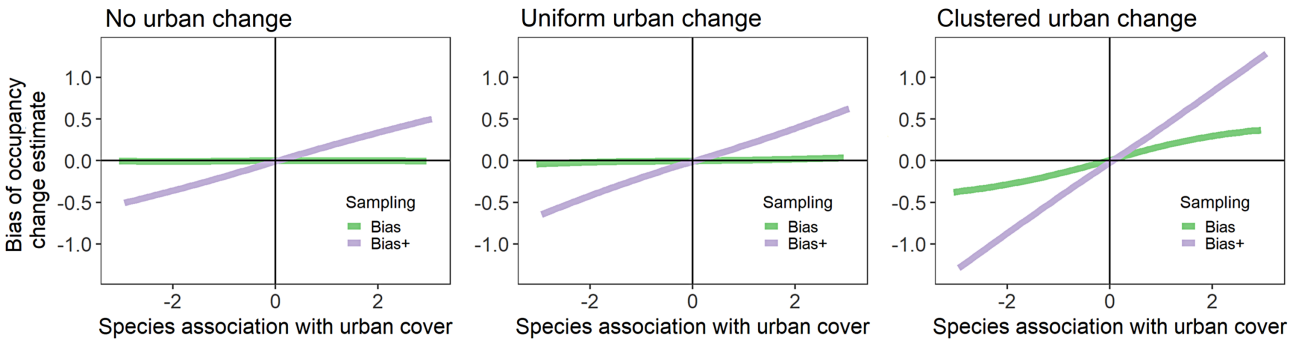

Figure 6. Effect on the bias of the occupancy change estimate of species habitat association with urban cover – occupancy of a species can have a negative association (urban avoider species), neutral association (i.e. association at zero – a generalist species) or positive association (urban exploiter species) with urban cover. Bias was assessed as the difference in the log-odds ratio of occupancy change between each sampling scenario (*bias* reflecting constant bias or *bias+* reflecting increasing bias) and the full sampling scenario. In all the previous main analyses and figures, a negative association was assumed.

have examined how species occurrence records are biased with respect to land cover or use change or biodiversity change drivers (Gonzalez et al. 2016, Shirey et al. 2021), or how spatial bias might have changed through time. In a recent study of butterflies in North America, the association between local human population density and sampling completeness (ratio between observed and expected species) became more strongly positive over time, based on observations in GBIF, iDigBio and eButterfly (Shirey et al. 2021). These findings are consistent with our own since human population density and urban cover are commonly strongly positively correlated. Taken together, it is likely that changing spatial bias of species occurrence records is widespread across different regions and databases.

Our example databases varied in data sources and hence different processes might be responsible for the changing spatial sampling bias in each. Different data sources of species occurrence records, including museum collections, scientific surveys, conservation monitoring and citizen science, are known to have different spatial biases and coverage (Geldmann et al. 2016, Speed et al. 2018, Petersen et al. 2021, Shirey et al. 2021). Moreover, citizen science itself is highly variable with some forms of citizen science associated with opportunistic presence-only records and other forms associated with robust data collected by coordinated projects with a standardized sampling protocol and spatial design (Isaac and Pocock 2015, Dobson et al. 2020). For heterogeneous databases, such as GBIF, increasing spatial bias over time may reflect an increased number of records from opportunistic citizen science data that are more prone to bias due to the lack of a coordinated sampling design (Boakes et al. 2010, Shirey et al. 2021). By contrast, within databases focused towards opportunistic or unstructured citizen science data, such as Naturgucker and Observation.org, increasing spatial bias might arise due to shifts in the types of citizen science projects and their participants. Barriers to participate in citizen science have lowered in the last decade due to new outreach projects and smartphone applications, leading to greater inclusion of people with lower expertise. Newer participants of citizen science may differ in their recording behaviour and be less likely to visit remote places for species observation compared with citizen scientists of earlier decades. In our analysis, we could not explicitly link increased bias to new participants but our test was limited by the variable quality of metadata on observers.

Implications of changing bias for biodiversity change research

Assessments of biodiversity change using species occurrence record databases have been criticized because of spatial biases in the underlying data (Cardinale 2014, Gonzalez et al. 2016, Fournier et al. 2019, Mentges et al. 2021). Our simulations showed how the effects of spatial sampling bias depend on a combination of changes in sampling bias through time, the pattern of environmental change, and the habitat associations of the species, which together determine whether drivers of species change are sampled representatively. For instance, oversampling of sites undergoing urbanization – either due to increases in the strength of spatial sampling bias through time or clustered environmental change with constant spatial sampling bias – leads to overestimation of declines of species that are negatively affected by urban cover. Since these effects increase with the strength of species' habitat associations, sampling biases are less likely to affect estimations of the trends of generalist species.

Several methods have been proposed to deal with the spatial bias of presence-only citizen science data in order to extract robust information of species' biodiversity patterns, but they mostly have assumed spatial sampling biases are constant over time. For instance, some studies have proposed the use of sampling weights, for instance, by upweighting undersampled areas, which has been more often applied to account for unrepresentativeness of sampling locations within standardized monitoring programs (McRae et al. 2017, Boersch-Supan et al. 2019, Johnston et al. 2020). The bias that we identify in our analysis can be regarded as a form of preferential sampling because the same environmental driver affects both species occupancy change and sampling locations. A common solution for preferential sampling involves modelling both the processes affecting species dynamics and those affecting site selection or visitation (Botella et al. 2021, Fandos et al. 2021). Pooling information from different data sources can often help separate species dynamics from the spatial biases in sampling (Dorazio 2014, Fithian et al. 2015, Pacifici et al. 2017). For instance, a promising new approach involves simultaneously modelling presence-only data along with standardized count or presence/absence data in so-called integrated distribution models (Dorazio 2014, Fithian et al. 2015, Pacifici et al. 2017).

Other studies have considered how spatial biases might be reduced by guiding the data collection of citizen scientists so that their sampling is more coordinated and representative (Callaghan et al. 2019a, b). For instance, heat maps could be used to visualize areas that are currently undersampled and hence where data collection would be especially useful for science. Since spatial bias is not the only issue that arises with opportunistic citizen science data (Altwegg and Nichols 2019), the value of opportunistic records could also be improved by more detailed metadata, e.g. on whether a set of observations reflect a complete checklist, enabling absence records to be inferred e.g. as used in eBird (Kelling et al. 2019), which is typically unknown in most species occurrence record databases. Also, promotion of more systematic biodiversity monitoring schemes with a proper sampling design, where reporting and sampling biases have been minimized, could serve to contribute more robust data but also play a pedagogical role in informing participating citizens about the importance of spatial sampling design. However, even coordinated citizen science projects can be affected by spatial sampling bias, due to differences in site-selection and retention rates with respect to land cover or use (Zhang et al. 2021).

Outlook

Large biodiversity databases have created huge opportunities for ecologists to ask questions about biodiversity patterns at large spatial scales (Theobald et al. 2015). The lack of standardized long-term monitoring for most taxa makes these data also especially valuable for the assessment of species change through time. At the same time, developments in statistical modelling have made it possible to account for many of the biases and sources of heterogeneity within heterogeneous data (Isaac et al. 2014). Increasing data bias towards urban areas might represent a challenge for some ecological questions but may also signal some opportunities. Large amounts of data within and nearby urban areas might provide natural experimental gradients to examine the impacts of future environmental scenarios, including climate warming (Lahr et al. 2018). Also, for promoting conservation awareness, increased data collection in urban areas may reflect an opportunity for greater engagement and reach across society of citizen science (Miller 2005) as well as increased value of urban green space. However, increasing amounts of data within urban areas may not fill data gaps in more remote areas that remain undersampled (Shirey et al. 2021). Hence, our results suggest that more intense efforts are needed to encourage data collection in a broader range of land covers and uses, near and remote from urban areas (Callaghan et al. 2019b). Until then, assessments of biodiversity change using heterogeneous databases should consider how temporal trends in spatial sampling bias might affect estimates of species long-term change.

Acknowledgements – This analysis was made possible through the efforts of many volunteer species recorders to whom we are very grateful. Open access funding enabled and organized by Projekt DEAL

Funding – We appreciate the support of the German Research Foundation (DFG) for funding the sMon working group (Trend analysis of biodiversity data in Germany) through the iDiv (DFG FZT 118, 202548816). CTC was also supported by a Marie Skłodowska-Curie Individual Fellowship (no. 891052).

Author contributions

Diana E. Bowler: Conceptualization (equal); Formal analysis (equal); Writing – original draft (equal). **Corey T. Callaghan**: Conceptualization (equal); Writing – review and editing (equal). **Netra Bhandari**: Visualization (equal); Writing – review and editing (equal). **Klaus Henle**: Conceptualization (equal); Writing – review and editing (equal). **M. Benjamin Barth**: Data curation (equal); Writing – review and editing (supporting). **Christian Koppitz**: Data curation (equal); Writing – review and editing (supporting). **Reinhard Klenke**: Conceptualization (equal); Writing – review and editing (equal). **Marten Winter**: Conceptualization (equal); Writing – review and editing (equal). **Florian Jansen**: Conceptualization (equal); Writing – review and editing (equal). **Helge Bruelheide**: Conceptualization (equal); Writing – review and editing (equal). **Aletta Bonn**: Conceptualization (equal); Funding acquisition (equal); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.06219>.

Data availability statement

Data is available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.4f4qrfjf4> (Bowler et al. 2022).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Altwegg, R. and Nichols, J. D. 2019. Occupancy models for citizen-science data. – Methods Ecol. Evol. 10: 8–21.

Amano, T. et al. 2016. Spatial gaps in Global Biodiversity Information and the role of citizen science. – Bioscience 66: 393–400.

August, T. et al. 2020. Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. – Sci. Rep. 10: 11009.

Bayraktarov, E. et al. 2019. Do big unstructured biodiversity data mean more knowledge? – Front. Ecol. Evol. 6: 239.

Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – Ecol. Inform. 19: 10–15.

Boakes, E. H. et al. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.

Boersch-Supan, P. H. et al. 2019. Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. – Biol. Conserv. 240: 108286.

Botella, C. et al. 2021. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. – Methods Ecol. Evol. 12: 933–945.

Bowler, D. E. et al. 2021. Winners and losers over 35 years of dragonfly and damselfly distributional change in Germany. – Divers. Distrib. 27: 1353–1366.

Bowler, D. E. et al. 2022. Data from: Temporal trends in the spatial bias of species occurrence records. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.4f4qrfjf4>.

Bradter, U. et al. 2018. Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species? – Methods Ecol. Evol. 9: 1667–1678.

Buckland, S. T. and Johnston, A. 2017. Monitoring the biodiversity of regions: key principles and possible pitfalls. – Biol. Conserv. 214: 23–34.

Burgess, H. K. et al. 2017. The science of citizen science: exploring barriers to use as a primary research tool. – Biol. Conserv. 208: 113–120.

Callaghan, C. T. et al. 2019a. Optimizing future biodiversity sampling by citizen scientists. – Proc. R. Soc. B 286: 20191487.

Callaghan, C. T. et al. 2019b. Improving big citizen science data: moving beyond haphazard sampling. – PLoS Biol. 17: e3000357.
Callaghan, C. T. et al. 2021a. Thermal flexibility and a generalist life history promote urban affinity in butterflies. – Global Change Biol. 27: 3532–3546.
Callaghan, C. T. et al. 2021b. Urban tolerance of birds changes throughout the full annual cycle. – J. Biogeogr. 48: 1503–1517.
Callaghan, C. T. et al. 2021c. Three frontiers for the future of biodiversity research using citizen science data. – Bioscience 71: 55–63.
Cardinale, B. 2014. Overlooked local biodiversity loss. – Science 344: 1098–1098.
Chandler, M. et al. 2017. Contribution of citizen science towards international biodiversity monitoring. – Biol. Conserv. 213: 280–294.
Collen, B. et al. 2008. The tropical biodiversity data gap: addressing disparity in global monitoring. – Trop. Conserv. Sci. 1: 75–88.
Dissanayake, R. B. et al. 2019. The value of long-term citizen science data for monitoring koala populations. – Sci. Rep. 9: 10037.
Dobson, A. D. M. et al. 2020. Making messy data work for conservation. – One Earth 2: 455–465.
Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. – Global Ecol. Biogeogr. 23: 1472–1484.
ESA 2017. Land Cover CCI product user guide ver. 2. Tech. Rep. – <maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf>.
Fandos, G. et al. 2021. Dynamic multistate occupancy modeling to evaluate population dynamics under a scenario of preferential sampling. – Ecosphere 12: e03469.
Feeley, K. J. and Silman, M. R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. – Divers. Distrib. 17: 1132–1140.
Fink, D. et al. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. – Ecol. Appl. 30: e02056.
Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.
Fournier, A. M. V. et al. 2019. Site-selection bias and apparent population declines in long-term studies. – Conserv. Biol. 33: 1370–1379.
Freeman, B. and Peterson, A. T. 2019. Completeness of digital accessible knowledge of the birds of western Africa: priorities for survey. – Condor 121: duz035.
Geldmann, J. et al. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. – Divers. Distrib. 22: 1139–1149.
Girardello, M. et al. 2018. Gaps in biodiversity occurrence information may hamper the achievement of international biodiversity targets: insights from a cross-taxon analysis. – Environ. Conserv. 45: 370–377.
Girardello, M. et al. 2019. Gaps in butterfly inventory data: a global analysis. – Biol. Conserv. 236: 289–295.
Gonzalez, A. et al. 2016. Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. – Ecology 97: 1949–1960.
Harrison, P. J. et al. 2014. Assessing trends in biodiversity over space and time using the example of British breeding birds. – J. Appl. Ecol. 51: 1650–1660.
Hugo, S. and Altwegg, R. 2017. The second Southern African Bird Atlas Project: causes and consequences of geographical sampling bias. – Ecol. Evol. 7: 6839–6849.
Husby, M. et al. 2021. Non-random sampling along rural–urban gradients may reduce reliability of multi-species farmland bird indicators and their trends. – Ibis 163: 579–592.
Isaac, N. J. B. et al. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. – Methods Ecol. Evol. 5: 1052–1060.
Isaac, N. J. B. and Pocock, M. J. O. 2015. Bias and information in biological records. – Biol. J. Linn. Soc. 115: 522–531.
Jimenez-Valverde, A. et al. 2019. Photo-sharing platforms key for characterising niche and distribution in poorly studied taxa. – Insect Conserv. Divers. 12: 389–403.
Johnston, A. et al. 2020. Estimating species distributions from spatially biased citizen science data. – Ecol. Model. 422: 108927.
Kelling, S. et al. 2015. Taking a 'Big Data' approach to data quality in a citizen science project. – Ambio 44: S601–S611.
Kelling, S. et al. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. – Bioscience 69: 170–179.
Lahr, E. C. et al. 2018. Getting ahead of the curve: cities as surrogates for global change. – Proc. R. Soc. B. 285: 20180643.
Mair, L. and Ruete, A. 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. – PLoS One 11: e0147796.
McRae, L. et al. 2017. The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. – PLoS One 12: e0169156.
Mentges, A. et al. 2021. Effects of site-selection bias on estimates of biodiversity change. – Conserv. Biol. 35: 688–698.
Meyer, C. et al. 2015. Global priorities for an effective information basis of biodiversity distributions. – Nat. Commun. 6: 8221.
Mihoub, J. B. et al. 2017. Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures. – Sci. Rep. 7: 41591.
Miller, J. R. 2005. Biodiversity conservation and the extinction of experience. – Trends Ecol. Evol. 20: 430–434.
Outhwaite, C. L. et al. 2020. Complex long-term biodiversity change among invertebrates, bryophytes and lichens. – Nat. Ecol. Evol. 4: 384–392.
Pacifici, K. et al. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. – Ecology 98: 840–850.
Petersen, T. K. et al. 2021. Species data for understanding biodiversity dynamics: the what, where and when of species occurrence data collection. – Ecol. Solut. Evid. 2: e12048.
Powney, G. D. et al. 2019. Widespread losses of pollinating insects in Britain. – Nat. Commun. 10: 1018.
Rousset, F. and Ferdy, J. B. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. – Ecography 37: 781–790.
Schmeller, D. S. et al. 2009. Advantages of volunteer-based biodiversity monitoring in Europe. – Conserv. Biol. 23: 307–316.
Sheard, J. K. et al. 2021. Long-term trends in the occupancy of ants revealed through use of multi-sourced datasets. – Biol. Lett. 17: 20210240.
Shirey, V. et al. 2021. A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. – Ecography 44: 537–547.

Speed, J. D. M. et al. 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. – PLoS One 13: e0196417.
Sullivan, B. L. et al. 2017. Using open access observational data for conservation action: a case study for birds. – Biol. Conserv. 208: 5–14.
Theobald, E. J. et al. 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. – Biol. Conserv. 181: 236–244.
Troudet, J. et al. 2017. Taxonomic bias in biodiversity data and societal preferences. – Sci. Rep. 7: 9132.
Yoccoz, N. G. et al. 2001. Monitoring of biological diversity in space and time. – Trends Ecol. Evol. 16: 446–453.
Zattara, E. E. and Aizen, M. A. 2021. Worldwide occurrence records suggest a global decline in bee species richness. – One Earth 4: 114–123.
Zhang, W. Y. et al. 2021. Habitat change and biased sampling influence estimation of diversity trends. – Curr. Biol. 31: 3656. e3–3662.e3.