Interpolation of temporal biodiversity change, loss, and gain across scales: a machine learning approach

Petr Keil [1] & Jonathan M. Chase [2, 3]*

[1] Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha – Suchdol, 165 00, Czech Republic

[2] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstrasse 4, 04103 Leipzig, Germany

[3] Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany

* Correspondence: pkeil@seznam.cz

Abstract

1. Estimates of temporal change of biodiversity, and its components loss and gain, are needed at local and geographical scales. However, we lack them because of data incompleteness, heterogeneity, and lack of temporal replication. Hence, we need a tool to integrate heterogeneous data and to account for their incompleteness.
2. We introduce spatiotemporal machine learning interpolation that can estimate cross-scale biodiversity change and its components. The approach naturally captures the expected and complex interactions between scale (grain), geography, data types, and drivers of change. As such it can integrate inventory data from reserves or countries with data from atlases and local survey plots. We present two flavors, both blending tree-based machine learning (random forests, boosted trees) with advances in ecological scaling: The first combines machine learning with species-area relationships (SAR method), the second with occupancy-area relationships (OAR method).
3. Using simulated data and an empirical example of global mammals and European plants, we show that tree-based machine learning effectively captures temporal biodiversity change, loss, and gain across a continuum of spatial grains. This can be done despite the lack of time series data (i.e., it does not require temporal replication at sites), temporal biases in the amount of data, and highly uneven sampling area. These estimates can be mapped at any desired spatial resolution.
4. In all, this is a user-friendly and computationally fast approach with minimal requirements on data format. It can integrate heterogeneous biodiversity data to obtain estimates of temporal biodiversity change, loss, and gain, that would otherwise be invisible in the raw data alone.

Keywords: biodiversity change, CART, extinction, invasion, MAUP, scale, time series

Introduction

Despite the concern that we face an unprecedented global alteration of biodiversity through extinctions, extirpations, invasions, and biotic homogenization, we still lack rigorous estimates of how fast, where, and at which scales this happens. The major problem with quantifying biodiversity change is lack of data. Only a small fraction Earth has been surveyed repeatedly in time, and there are geographic gaps in most taxa and regions (Meyer et al., 2015; Mora et al., 2008); for example, even the most comprehensive database of local biodiversity change (Dornelas et al., 2018) misses many taxa and vast geographic regions (i.e., little coverage in much of South America, Africa, and Asia). Given the cost of biodiversity surveys repeated in time and over large scales, this data deficiency will likely persist.

Apart from the data problems, it has been increasingly clear that analyses of biodiversity trends need to acknowledge the grain dependency of biodiversity change (Chase et al., 2019) and its components: loss (Keil et al., 2018) and gain (Englund & Hambäck, 2007). The reason is that biodiversity change at the global scale can be decoupled from local or regional change, and vice versa (Fig. 1) (Chase et al., 2019; Keil et al., 2018). For example, even though species richness is declining globally via extinction (Barnosky et al., 2011), local trends show increases, decreases and stasis, often with an average of no net change (Blowes et al., 2019; Dornelas et al., 2014; Vellend et al., 2013). Yet, integrative studies mapping diversity change across grains are lacking. Instead, some report extinction rates at the global level and observed over centuries (Alroy, 2015; Barnosky et al., 2011), others examine biodiversity change in local assemblages over years or decades (Blowes et al., 2019), while rates of invasion are often studied at the regional level (e.g., a country or state level) (van Kleunen et al., 2015). Further, data from different grains suffer their own grain-specific deficiencies and challenges. For instance, fine-grain time series are geographically clumped (Dornelas et al., 2018); in contrast, although less clumped, coarse-grain data on extinctions (https://www.nationalredlist.org/) and invasions (van Kleunen et al., 2019) are only available for political administrative units that are highly variable in their area, hindering direct comparisons across grains.

Here, we present a method that addresses these issues. It is based on the idea that, instead of collecting more data, we can learn from the strengths of different data types at different grains, and interpolate biodiversity change into unsurveyed areas and spatial grains. In (Keil & Chase, 2019), we used a similar idea to integrate heterogeneous data from local plots and regional checklists to estimate *static* patterns of species richness, and their environmental drivers, across a continuum of spatial grains. Here we generalize the method to include time. In short, using tree-based machine learning (Breiman et al., 1984; Hastie et al., 2011; Viana et al., 2019), we model species richness and species occupancy as a function of the interaction between geographic coordinates, area, and time. The method can use heterogeneous data to reliably estimate temporal change of species richness, and its components loss and gain, across spatial grains, and can map the estimated biodiversity trends geographically. This can be done even when there are no repeated estimates of biodiversity in one location through time (i.e., time series).

Material and Methods

Theoretical background

Our proposed approach stands on four principles:

Principle 1: Grain dependence of diversity change can be modelled as an interaction between time and area. First, a grain dependent effect of time on biodiversity can be assessed from a statistical interaction between time *t* and area A of spatial units (sites, regions, polygons). When the A is constant across a set of spatial units, we call it *grain*. Let's consider change of species richness as a function of time and area in power-law *species-area relationship* (SAR):

$$\hat{S}_i = \beta_0 A_i^{z_{t_i}} \quad \text{(eq. 1)}$$

where $\hat{S}_i$ is expected mean number of species at *i*-th observation (data point) of area A_i and time t_i. β_0 is constant, and z_{t_i} is the time-dependent SAR exponent. We can assume that z changes as a linear function of *t* as $z_{t_i} = \beta_1 + \beta_2 t_i$, and thus $\hat{S}_i = \beta_0 A_i^{(\beta_1 + \beta_2 t_i)}$, which is:

$$\ln(\hat{S}_i) = \ln(\beta_0) + \beta_1 \ln(A_i) + \beta_2 \ln(A_i)\, t_i \quad \text{(eq. 2)}$$

We can further add an area-independent effect of time $\beta_3 t_i$:

$$\ln(\hat{S}_i) = \ln(\beta_0) + \beta_1 \ln(A_i) + \beta_2 \ln(A_i)\, t_i + \beta_3 t_i. \quad \text{(eq. 3)}$$

The observed species richness S_i can be modelled, for example, as $S_i \sim Poisson(\hat{S}_i)$, which together with eq. 3 give us a Poisson generalized linear model (GLM) with log link function and with an interaction term between time and area. This GLM can be trained on data on species richness across spatial units of varying area, and then used to predict $\hat{S}$ at any chosen location of an arbitrarily chosen area and time, as long as these fall within the range of values of the training data. The predicted temporal change of richness at a given area between times 1 and 2 can be then expressed as a simple difference,

$$\hat{S}_\Delta = \hat{S}_{t=1} - \hat{S}_{t=2}, \quad \text{(eq. 4)}$$

or as a log ratio:

$$LR = \log_{10}(\hat{S}_{t=1}/\hat{S}_{t=2}). \quad \text{(eq. 5)}$$

The interaction between time and area in eqs. 2 and 3 allows for the seeming paradox of $\hat{S}_\Delta$ or LR being positive at one grain and negative at another, as observed by several studies (Chase et al., 2019; Keil et al., 2011; Powell et al., 2013).

Principle 2: Occupancy-area relationship (OAR) is a single-species version of SAR. In eqs. 1-5 we only focused on species richness. But the approach can be applied to individual species, and thus to interpolate species *composition* when, for each location, species identities are available.

The relationship between area and probability of observing a single species is known as the occupancy-area relationship (OAR) (Azaele et al., 2012; Kunin, 1998), which can be understood as a single-species SAR. Let P_{ij} be the probability of observing a species *j* at site *i* of

area A_i at time t_i, where $j \in \{1,2,\ldots,S_{tot}\}$, and S_{tot} is the total number of species in the entire extent of our study. If we consider no effect of spatial coordinates, then P_{ij} is the frequency of species across all sites of a given area at time t_i, which is sometimes termed *occupancy* (Azaele et al., 2012). Analogically to the time-dependent SAR in q. 2, we can write a time-dependent OAR as:

$$logit(P_{ij}) = \ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \ln(\beta_{j0}) + \beta_{j1}\ln(A_i) + \beta_{j2}\ln(A_i)\,t_i \quad \text{(eq. 6)}$$

The observed binary presence or absence O_{ij} is then $O_{ij} \sim Bernoulli(P_{ij})$. Note that since each of the β_j coefficients is species-specific, we effectively have a total number of S_{tot} separate GLMs, which means that there is a statistical *interaction* between species identity and the effects of area and time. The GLM in eq. 8 can be further extended by adding a random effect of species identity, non-linear effects of time and area, and/or by including geographic coordinates and the interaction terms mentioned above. Once fitted, this GLM can predict P_j at any chosen location of an arbitrarily chosen area and time. From the predicted P_j, we can then get to expected species richness as:

$$\hat{S} = \sum_{j=1}^{S_{tot}} P_j \quad \text{(eq. 7)}$$

Moreover, for the *j*-th species and at a given area, we can predict probability of extinction between time 1 and 2 as $P_{ext_j}(A) = P_{j,t=1}(1 - P_{j,t=2})$ and probability of species gain as $P_{gain_j} = P_{j,t=2}(1 - P_{j,t=1})$. From these we can get to the total expected number of extinct species at a given area as:

$$\hat{S}_{ext} = \sum_{j=1}^{S_{tot}} P_{ext_j}, \quad \text{(eq. 8)}$$

and gained species,

$$\hat{S}_{gain} = \sum_{j=1}^{S_{tot}} P_{gain_j}, \quad \text{(eq. 9)}$$

and thus

$$\hat{S}_{\Delta} = \hat{S}_{gain} - \hat{S}_{ext} \quad \text{(eq. 10)}$$

Further, predicted P_j values for multiple species can be used to calculate species temporal turnover at a site, or temporal change of spatial turnover between sites.

Principle 3: Diversity and its drivers follow gradients which can be interpolated. Species distributions are spatially aggregated, which is also reflected in spatial autocorrelation of species distributions and richness (Dormann et al., 2007; McGill, 2010). Because of this, richness at a given grain and time tends to follow spatial gradients (Lomolino et al., 2010), where a gradient is defined as a systematic change of richness along one or more spatial coordinates. An example is latitudinal gradient of richness: $\frac{d\ln(\hat{S})}{dLat} = f(Lat)$. Furthermore, not only richness, but also the *effects* (coefficients) of other variables on richness follow spatial gradients; for example the ratio of regional and local diversity (which directly relates to coefficient z of SAR, eq. 1) in Palearctic region has been shown to increase towards South-East Asia (Keil & Chase, 2019). It has also been shown that effects of environmental predictors on richness follow geographic gradients (Whittaker et al., 2006). To represent this mathematically, instead of the spatially constant (stationary) coefficient β_3 from eq. 3 consider β_3 as a

function $\beta_3(Lat)$ which follows a latitudinal gradient, i.e. $\frac{d\beta_3(Lat)}{dLat} = g(Lat)$. When $g(Lat)$ is a constant, then we have a linear gradient, and by integration we get a statistical *interaction* between time and latitude $const. + \beta_3 t_i Lat_i$, instead of $\beta_3 t_i$ in eq. 3. This logic can be extended to any term in eq. 3 to get both independent effect of predictors (A, t, Lat) on species richness as well as their interactions:

$$\ln(\hat{S}_i) = \ln(\beta_0) + \beta_1 \ln(A_i) + \beta_2 \ln(A_i)\, t_i + \beta_3 t_i + \beta_4 \ln(Lat_i) + \beta_5 t_i \ln(Lat_i) + \beta_6 \ln(A_i)\, Lat_i + \beta_7 \ln(A_i)\, t_i Lat_i \quad \text{(eq. 11)}$$

Eq. 11 is an example of a model that can be used to interpolate richness along spatial coordinate(s) into locations and grains with no data. When eq. 5 is fitted to data with varying A and t of each spatial unit, it can predict $\hat{S}$, $\hat{S}_\Delta$, or LR continuously along latitude, and at grains and locations that have not been surveyed. Because of the interaction terms, eq. 5 allows $\hat{S}_\Delta$ and LR to have different magnitudes (and directions) at different parts of the geographic gradient and at different grains.

Principle 4: Tree-based machine learning suits complex interactions between area, time, and their non-linear effects. So far, we have considered simple gradients along one spatial dimension. However, data on real-world biodiversity are 2- or 3-dimensional and diversity follows complex and non-linear geographic gradients (Lomolino et al., 2010). Further, SARs are unlikely to follow a simple power law over large spans of grains (Storch, 2016), and temporal trends of richness can be non-linear, for example, when responding to abrupt anthropogenic pressures (Jung et al., 2019). Because of these complexities, generalized linear models (GLM) introduced above, and related parametric methods, are prone to over-simplification and misspecification. We propose that this can be addressed by using flexible machine learning algorithms based on *classification and regression trees* (CART), and methods derived from CART such as *random forests* (RF) or *boosted regression trees* (BRT) (Breiman et al., 1984; Hastie et al., 2011), hereafter we will use *tree-based methods* for all of these.

Statistical interactions and non-linear responses are implicit in the very construction of CART, and their shape is estimated from the data without requiring the researcher to have any preconception on this matter (Schiltz et al., 2018). Furthermore, the logic described in eqs. 1-5 applies equally to GLM and CART, because the expected value in both approaches follow the same general formula $\boldsymbol{X\beta}$, where matrix $\boldsymbol{X}$ is matrix of predictors and $\boldsymbol{\beta}$ is the vector of coefficients. For example, eq. 1 describes the expected value of species richness ($\hat{S}$) in a Poisson GLM, which can be re-written using matrix notation as $\widehat{\boldsymbol{S}} = \boldsymbol{X\beta}$, which reads: the vector of expected values of species richness ($\widehat{\boldsymbol{S}}$) is a product of the matrix of predictors $\boldsymbol{X}$ and vector of coefficients $\boldsymbol{\beta}$. This can be expanded as:

$$\begin{pmatrix} \ln(\hat{S}_1) \\ \vdots \\ \ln(\hat{S}_n) \end{pmatrix} = X\beta = \begin{pmatrix} 1 & \ln(A_1) & t_1 & \ln(A_1)\, t_1 \\ 1 & \vdots & \vdots & \vdots \\ 1 & \ln(A_n) & t_n & \ln(A_n)\, t_n \end{pmatrix} \beta. \quad \text{(eq. 12)}$$

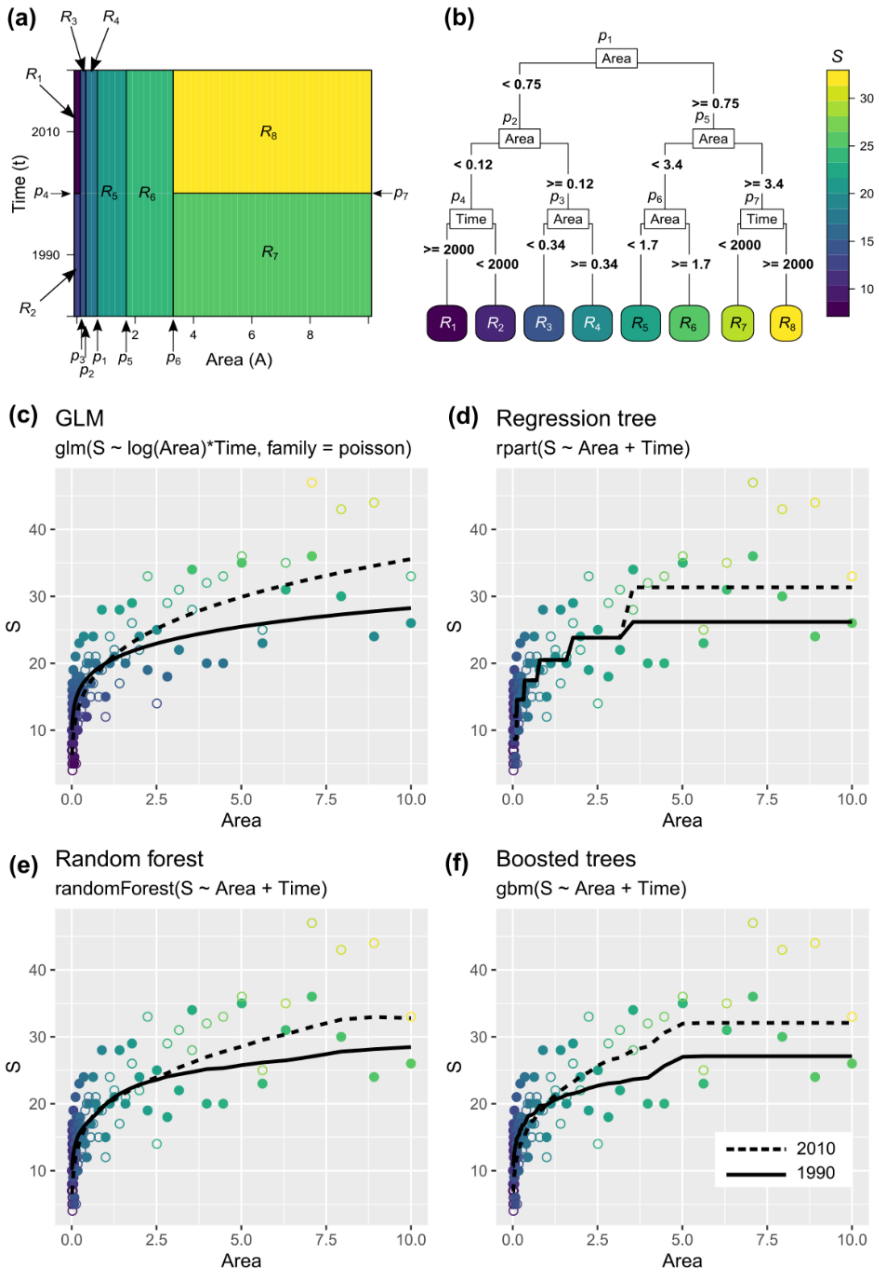

Figure 1 *Alternative ways to model mean species richness* $\hat{S}$ *as a function of area and time. (a-b) Illustration of a regression tree algorithm in which the cartesian space defined by the area and time is divided into* K *regions* R_k *at splitting points located at values* p_l*. (c-f)* $\hat{S}$*, represented by black lines fitted through observed richness S (points) using four alternative algorithms, called by their respective* R *functions.*

In regression trees, as in GLM, the expected value is $\widehat{\boldsymbol{S}} = \boldsymbol{X\beta}$. However, instead of using the predictors such as A and t directly, the cartesian space defined by the predictors is divided into K regions R_k at splitting points located at values t_p along the predictors (Fig. 1a), and these regions are represented by binary vectors. Regions R_k are identified using a recursive

splitting algorithm that, at each split t_p, maximizes the difference between expectancy in the two groups resulting from the split. Thus, we get:

$$\begin{pmatrix} \ln(\hat{S}_1) \\ \vdots \\ \ln(\hat{S}_n) \end{pmatrix} = \boldsymbol{X\beta} = \begin{pmatrix} 1 & R_{11} & \dots & R_{1K} \\ 1 & \vdots & \cdots & \vdots \\ 1 & R_{n1} & \dots & R_{nK} \end{pmatrix} \boldsymbol{\beta}. \qquad \text{(eq. 13)}$$

In Fig. 1 we see that a simple regression captures both the non-linearity and the time-area interaction even though these are not explicitly specified in the function call. This is an advantage over the GLM, where these complexities need to be specified á priori. However, the downside of a single tree is that it only proceeds in discrete steps, and it can be prone to overfitting. Fortunately, RF and BRT address this problem (Fig. 1), producing smooth non-linear responses, while avoiding overfitting and offering the same flexibility and simplicity as a regression tree.

Practical recipe

We have shown that species richness and individual species' probability of occurrence can be modelled as a function of an interaction between area, time, and spatial coordinates, and these models can be used to predict gain, loss, and change of biodiversity at any grain, and interpolated to locations and times with no data. This can be done using GLM, or more conveniently, using conceptually similar but more flexible tree-based machine learning. Here we show how to do this in R. We propose 2 approaches, depending on the type of data at hand:

(i) **SAR method** (Fig. 2 left). When we only have data on species richness of each location, we can fit the time-dependent species-area relationship (eqs. 1-3), or its extension with spatial coordinates (eq. 11).

(ii) **OAR method** (Fig. 2 right). When we also have species composition for each location, we can use the time- and species-dependent occupancy-area relationship (eq. 6), which we fit for all species (together, in one model), and we then summarize the predicted P_{ij} of each species to any desired quantity (eqs. 7-10).

First, we start with the data - these can be from local plots, nature reserves, counties, or countries (hereafter *polygons*, Fig. 2a). We need information either on species richness or species composition (presences/absences) of each polygon. The polygons can vary considerably in their area and each individual polygon can surveyed once, or more than once. Even though Figure 2a considers two temporal periods (for the sake of simplicity), each polygon can be surveyed at a different time.

The second step (Fig. 2b) is to format the data. For the SAR method, we need a data frame with *n* rows (*n* is the number of observations) with observed species richness (S_{it}), survey time (date, year), polygon area, and geographic coordinates. For the OAR method we need a data frame of nS_{tot} rows, and instead of richness, we need data on occupancy (binary presence/absence O_{ij}) of every species, and an additional categorical variable identifying each species.

Third, we fit a GLM, or a machine learning algorithm that allows for complex interactions and non-linear effects (RF or BRT are particularly suitable), with either S_i or O_{ij} as a response, and with time, area, geographic coordinates, and categorical species identity, as predictors.

Finally, we can use the fitted object to predict several features of biodiversity change (Fig. 2ed-f). For example, we can predict static SAR or OAR relationships at any desired time period (Fig. 2d), we can calculate average $\hat{S}_{\Delta}$ (or the separate components of species gain and loss) across the whole spatial extent at any grain (Fig. 2e) using eqs. 8 and 9, and we can predict spatially explicit and contiguous maps of biodiversity change at any grain (Fig. 2f).

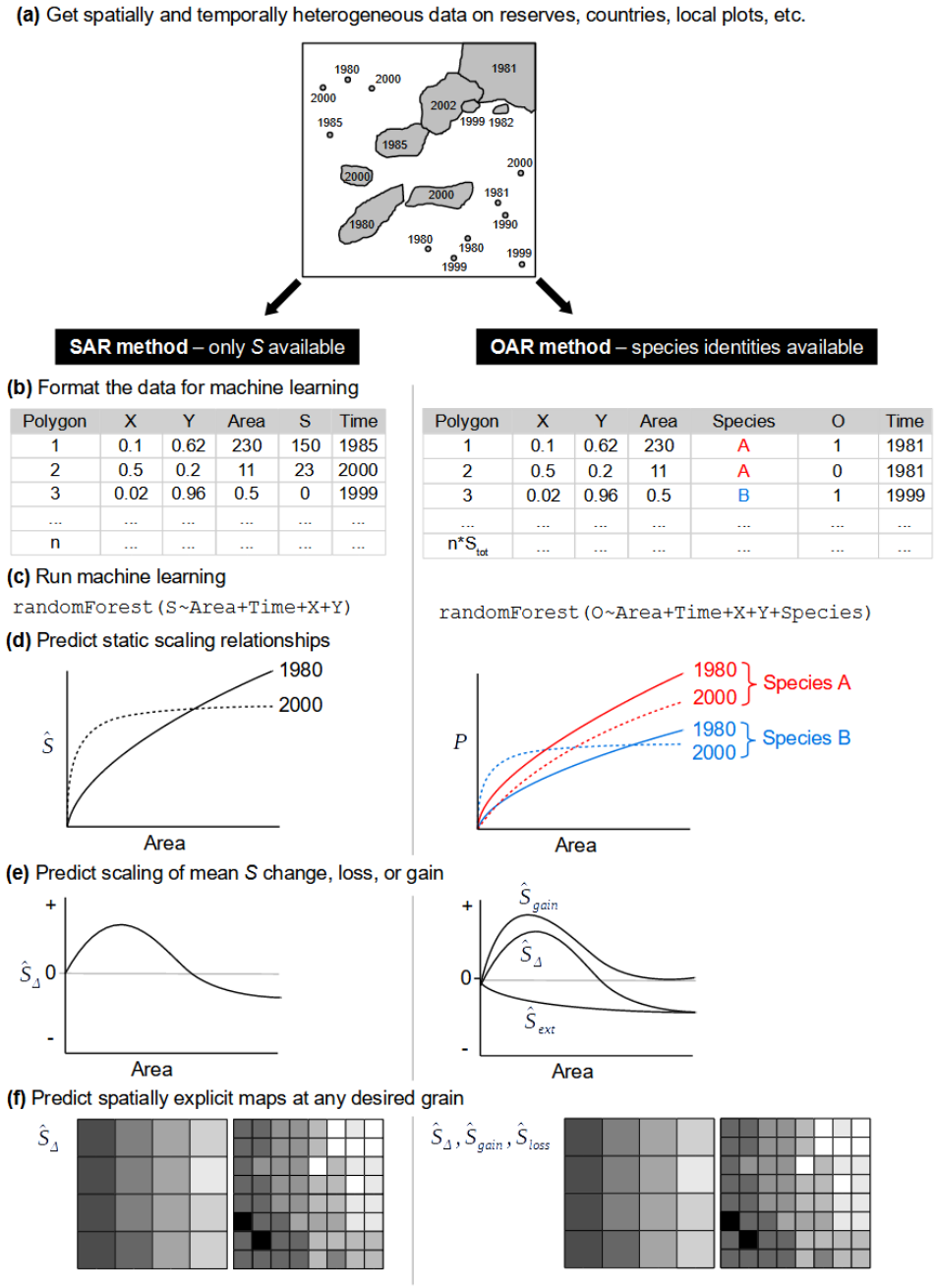

Figure 2 *Illustration of the practical workflow of applying the SAR-method or the OAR-method on heterogeneous polygon data varying in their area, time, and lacking temporal replication.*

Test on simulated data

We tested the performance of the SAR and OAR methods described above on simulations of grain-dependent biodiversity change. In these simulations we also examined the sensitivity of the approach to sampling effort, temporal bias of sampling effort, as well as sensitivity to variation in the total number of species, magnitude of gain and loss, and spatial grain of the predictions. All the code for the simulations and their analyses is at https://github.com/petrkeil/S_change_interpolation.

Simulations. Our aim was to simulate situations with potentially different directions of biodiversity change at different grains. In each simulation, we first generated a community of S_{tot} species at time 1, each species as a spatially explicit point pattern within a square domain of side of length 1, with a S_{abs} fraction of S_{tot} being absent from the community at time 1. Individuals (i.e. points) of each species *j* were aggregated in N_{clust_j} spatial clusters, where N_{clust_j} values were sampled from a log-normal distribution, and rounded to whole numbers. On average, every species had N_{ind_j} individuals per cluster, with the total number of individuals per species being $N_j = N_{ind_j} N_{clust_j}$. We then generated a community at time 2 by subjecting each species in community at time 1 to random loss or gain of individuals or clusters, or both. The loss and gain at the level of clusters and individuals enabled the grain-dependent biodiversity change. This was modulated by three parameters: $P_j(I|N_j = 0)$, the probability of immigration *I*, i.e. the probability of an absent species to re-appear; $-1 < \Delta_{clust} < 1$, the temporal trend of N_{clust_j}; $-1 < \Delta_{ind} < 1$, the temporal trend of N_{ind_j}. We simulated this procedure for all combinations of the following parameter values: $S_{tot} \in \{13, 26, 52\}$, $N_{clust_j} \in \{100, 500\}$, $N_{ind_j} \in \{100, 500\}$, $P_j(I|N_j = 0) \in \{0, 0.2, 0.7\}$, $\Delta_{clust} \in \{-0.7, -0.3, 0, 0.3, 0.7\}$, and $\Delta_{ind} \in \{-0.7, -0.3, 0, 0.3, 0.7\}$, resulting in 16,200 simulations in total. For each simulation we then calculated $\hat{S}_{\Delta}$ at four spatial grains with areas of cells of 0.001, 0.0039, 0.156 and 0.0625 (corresponding to grid cell side lengths of 32, 16, 8, and 4).

Sampling. For each simulation we generated a set of non-overlapping polygons of varying area, representing real-world shapes such as reserves, or administrative regions (Fig. 2a). We generated the polygons like this: Within the community domain we simulated a realization of the Thomas cluster process (Diggle et al. 1976) with a constant intensity 5, random displacement standard deviation from cluster centers of 0.08, and mean number of points per cluster of 50. We used these points as centers of Dirichlet tesselation in R package spatstat (Baddeley et al., 2016).

We then sampled fractions P1 and P2 of the total number of polygons in time 1 and 2 respectively, with P1 and P2 having all combinations of 0.2 and 0.4, so that we had temporally balanced ($P1 = P2$) and temporally biased sampling ($P1 \neq P2$), as well as lower ($P1 = P2 = 0.2$) and higher ($P1 = P2 = 0.4$) sample size. None of the polygons was surveyed twice in time. We then extracted species composition, species richness, area, and geographic coordinates of each polygon, and we also extracted these values for grid cells in regular grids of $A \in \{0.001, 0.0039, 0.0156, 0.0625\}$, corresponding to grid cell side lengths of 32, 16, 8, and 4. These then represented our "true" values that we then aimed at predicting with the machine learning approaches.

Analyses. We fitted four variants of the SAR method and one variant of the OAR method to the polygon data (these correspond to columns in Fig. 3):

1. SAR-based, with only A and t as predictors (Fig. 3a).
2. SAR-based, with A, t, and X and Y spatial coordinates as predictors (Fig. 3b).
3. SAR-based, fitted separately in each time, with A, X and Y as predictors (Fig. 3c).
4. SAR-based, fitted separately in each time, with only A as predictor (Fig. 3d).
5. OAR-based, with A, t, X and Y as predictors (Fig. 3e)

For each of these variants we used random forests using R package `randomForest` and boosted regression trees using R package `gbm`. These were used to predict average $\bar{S}_{it}$ at the four spatial grains described above, and we compared the predicted values with the simulated "truth".

Results. Using the simulations, we found that the all versions of the SAR-based method, as well as the OAR method, estimated average species richness change $\Delta\bar{S}_i$ that was in line with the true change (Fig. 3). Algorithms without spatial coordinates X and Y performed worse than those with them, and the OAR method gave slightly better correlation between true and predicted values than SAR methods (Fig. 3). When time t was not a predictor in the models, but the two time periods were modelled separately, the true-predicted correlation was lower than when t was a predictor (Fig. 3). We did not find a clear indication that either RF or BRT performed better than the other (Fig. 3). Focusing specifically on the SAR method and the `S ~ A + t + X + Y` formula implemented in BRT, we found that predictive errors decreased towards fine grain (Fig. S1), decreased with increasing sampling effort in the less-surveyed time period (Fig. S2), but were not as sensitive to temporal bias (Fig. S3). We didn't find any clear effect of grain, sampling effort, or temporal bias on the prediction bias (Figs. S1-S3).

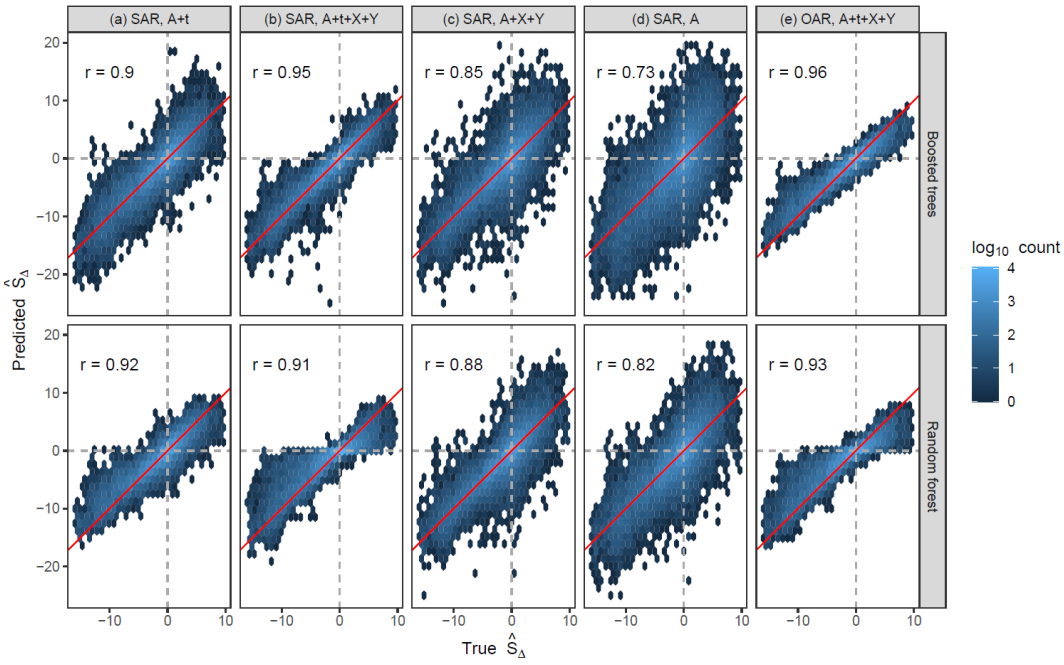

Figure 3. *Average $\hat{S}_\Delta$ predicted by tree-based machine learning trained on messy polygon data with no temporal replication versus the "true" $\hat{S}_\Delta$ from the simulations. Rows indicate the two algorithms (random forest or boosted trees). The first four columns are variants of the SAR-based method, the fifth column is the*

OAR method. Red diagonals are 1:1 lines, r is Pearson correlation coefficient, A is area, t is time, X and Y are spatial coordinates. Shades of blue indicate count of simulations.

Application to empirical data

Data and code necessary to reproduce the empirical analyses is openly available at: https://github.com/petrkeil/S_change_interpolation.

Global mammals. We used the Phylacine 1.2.1 atlas (Faurby et al., 2018) of mammal macroecology for the first illustration of our approach. Phylacine provides two types of geographic ranges of 5,831 known mammal species: (1) *current ranges*, which are the present-day distributions over 1-degree global grid (Behrmann projection), and (2) *present natural ranges*, which are assumed ranges that a species would have if it had never experience anthropogenic pressures. This category also encompasses species that went extinct during the Holocene, such as mammoths, Coelodonts, Smilogons, or ground sloths, as well as species that underwent range expansion. Here we treat the present natural ranges to represent Time 1, and the current ranges to represent Time 2, in line with (Faurby & Svenning, 2015). We excluded all marine species and all species ≤ 20 kg from the database, assuming that large species are better mapped and preserved in the fossil record, leaving us with 444 species.

We then used the geographic ranges to create a spatially and temporally incomplete dataset for the SAR and OAR methods. Specifically, we first overlaid the range maps with polygons of countries, using Level 3 classification (Fig. 4a) of Brummitt (2001). We randomly assigned each country polygon as being sampled either in Time 1 or Time 2. For each polygon we noted all mammal species, and species richness, that could be detected within its boundaries in the given time. Thus, each polygon has only been sampled once, either in time 1 or 2. We created a hexagonal grid with grid cells of area of 209,903 km² (Fig. 4b) from which we calculated the true $\hat{S}_{ext}$, $\hat{S}_{gain}$, $\hat{S}_{\Delta}$ and *LR* directly from the Phylacine database. We then compared these "true" values with the values estimated by the machine learning approach from the incomplete dataset. We used the boosted regression trees algorithm from R package `gbm` for both the SAR method and the OAR method (Fig. 2).

Looking at the Holocene empirical change in 444 species of mammals heavier than 20 kg, we found that both the SAR and OAR methods predicted geographic patterns of $\Delta\bar{S}_i$ derived from the country-level incomplete data matched the true patterns (Fig. 4b-d), although the true vs predicted match was considerably better for the countries (see the online repository) then for the regular hexagons (Fig 4). Further, the OAR method gave unbiased predictions of $\hat{S}_{\Delta}$ and $\hat{S}_{ext}$ outperformed the SAR method in estimation of $\Delta\bar{S}_i$, but over-estimated $\hat{S}_{gain}$ (Fig. 5b).

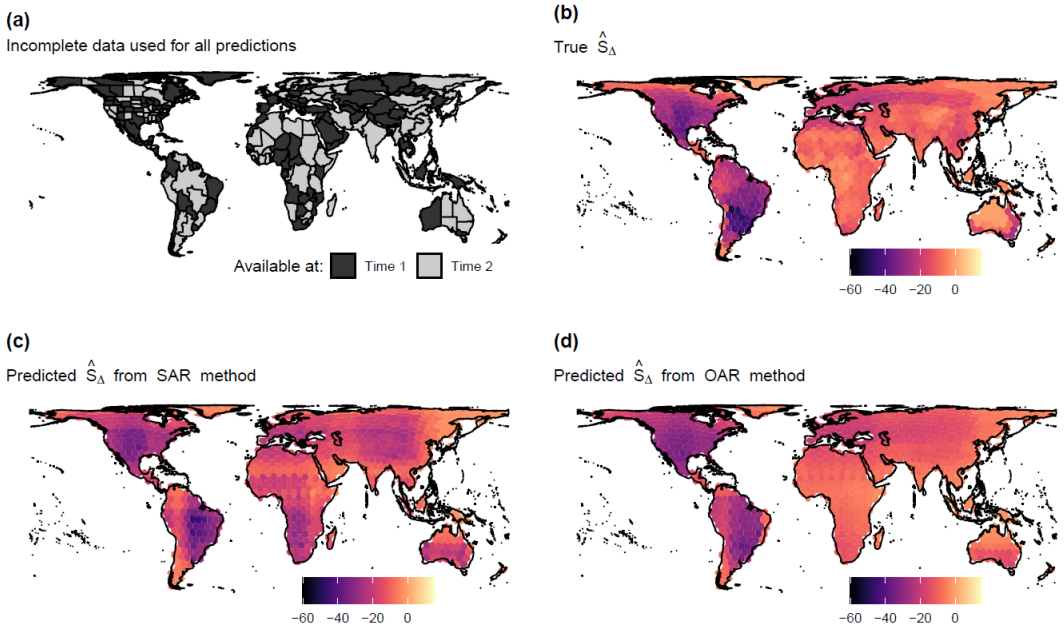

Figure 4. *Maps of true change of species richness of large mammals ($\hat{S}_\Delta$, panel b), and $\hat{S}_\Delta$ predicted by boosted regression trees (panels c and d) trained on spatially heterogeneous and temporally incomplete data (panel a). We stress that each of the countries in panel only had data from a single temporal period, and the countries varied in shape and area. Yet, the interpolation algorithm could predict temporal change of diversity in a regular grid that correlated well with the true change (see Fig. 5).*

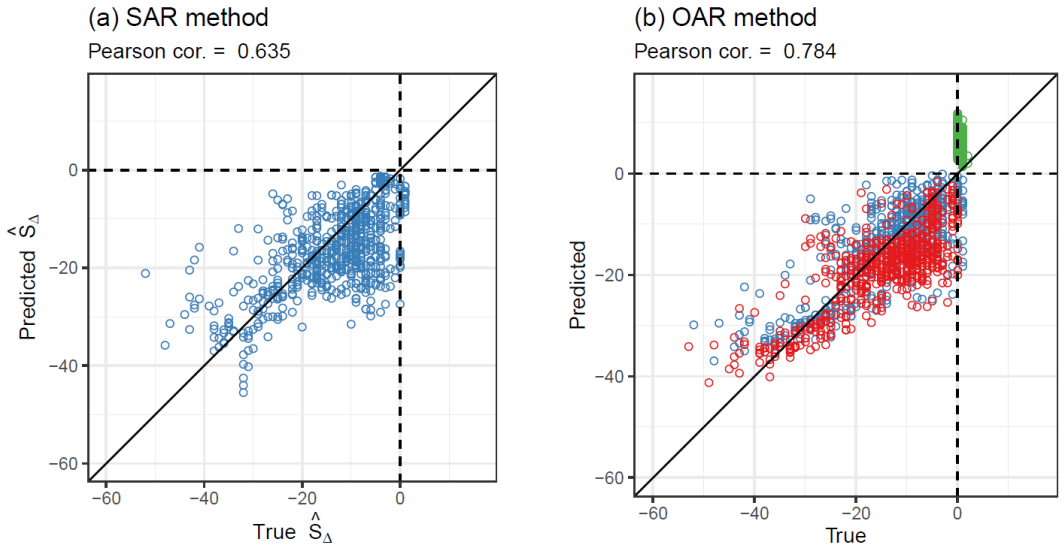

Figure 5. *Comparison of true values of mammal biodiversity change ($\hat{S}_\Delta$, blue; $\hat{S}_{ext}$, red; $\hat{S}_{gain}$, green) with values predicted by boosted regression trees trained on spatially heterogeneous and temporally incomplete data. Each point here represents a hexagon of equal area, as presented in Fig. 4b-d.*

European plans. We also tested our machine learning approach on European plants. Specifically, we used data by Essl *et al.* (2013), who provide data on all non-native, native, and extinct plant in individual European countries (Fig. 6a). In this case our aim was to use the SAR method (Fig. 2) to standardize the observed change of species richness to a regular grid, and to look for a pattern that might have been obscured by the varying area of European countries. To do this, we used the `randomForest` function in R package randomForest (Breiman et al., 1984). After predicting the log ratios of species richness to the regular grid, we found a latitudinal gradient of species gain, with higher gains towards the north (Fig. 6b), which is something that would otherwise be obscured in the country-wide data. Further, the magnitude species gain in the cells of the grid is considerably lower than the gain on the level of northern European countries.

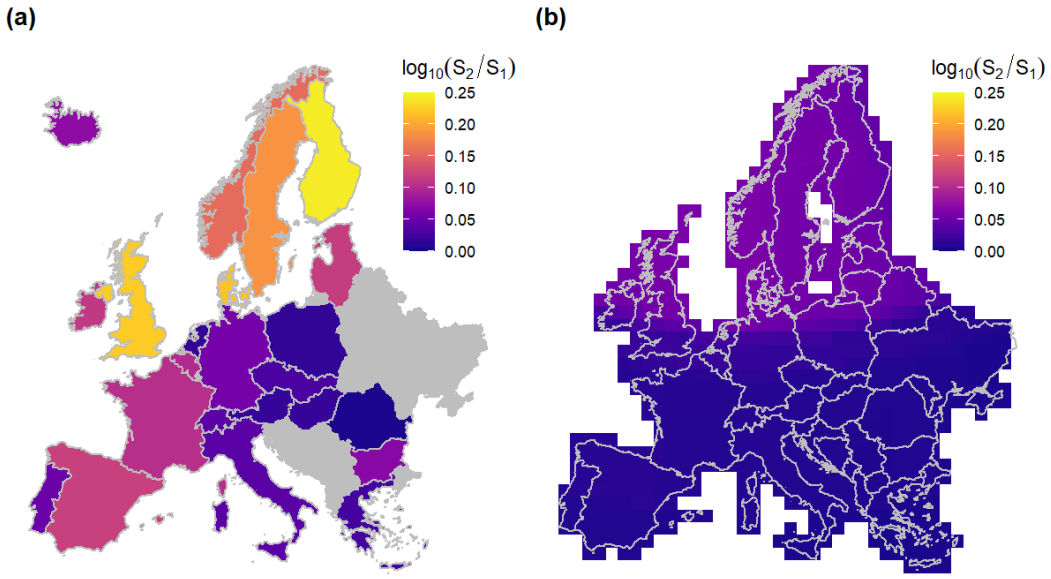

Figure 6. *Maps of change of species richness (expressed as log ratio) of European plants. Panel (a) shows raw data from Essl et al. (2013), grey areas have no data available. Panel (b) is species richness change predicted by a random forest-based SAR method, with the raw data as an input.*

Discussion

Overall, our machine learning approach produced mostly unbiased estimates of temporal biodiversity change in both simulated data and in empirical data on mammals. This worked even when temporally replicated data were unavailable, and when sampling effort was uneven in time. On the example of European plants, we have further demonstrated that the approach can uncover spatial patterns of change that may not be obvious in heterogeneous raw data. Thus, even though there is a margin for improvement (see below), our approach can be a foundation for larger data integration efforts, and particularly promising for application on messy data that vary in their area, and that lack temporal replication. This data deficiency is the case of most regions outside the wealthy global North (Meyer et al., 2015), particularly in the tropics. In these regions, it is unlikely that we will soon have high-quality

large-scale standardized biodiversity data. Thus, if we want to get an idea about how biodiversity changes in time in these regions, standardization and interpolation methods will be essential. Here we offer a user-friendly take on such a method.

Sensitivity to biases in data. In the simulations, our approach underperformed when the sampling effort was low, i.e. when the input data have low *N*, for example at the coarsest spatial grains (Fig. S1, S2). This can also explain the weak performance when predicting species gain in the empirical dataset (Fig. 5), since there was only a small number of large mammal species that expanded their geographic ranges. We thus advise caution when interpolating (or extrapolating) to regions or grains with low sample sizes, unless the uncertainty due to low sample size is taken into account (e.g. using Bayesian inference, as in Keil & Chase, 2019). Encouragingly, however, the approach worked well in the face of temporal bias in the amount of data (Fig. S3); this will often be the case in real-world data, particularly when comparing the pre- and post-2000 periods, where the availability of GPS and mobile computing devices dramatically increased the amounts of data.

Possible improvements. Indeed, we only present the approach in its simplest form, and additional predictors can easily be added, which include the shape and elongation of polygons which are known to influence diversity estimates (Kunin, 1997), variables capturing sampling methodology (e.g. as in Blowes *et al.*, 2019), variables representing drivers of biodiversity and its change, such as climate or land use (as in Keil & Chase 2019), and a variable representing the known effect of temporal grain (Foote, 1994). Another simplification is that so far we focused on prediction of the expected mean value at a given location, tree-based machine learning can be used for both inference about the effects of predictors, as well as calculation of statistical significance of effects of predictors (R package `ranger`, Janitza *et al.*, 2018; R package `randomForestCI`, Wagner *et al.*, 2014). Also, R packages `randomForest` and `gbm` offer implementations to visualize marginal effects of predictors through `patialPlot` and `plot.gbm` respectively, and both can extract relative importance of both individual predictors and their interactions. For an overview of recent advances in ML-based inference in ecology see Lucas (2020). Clearly, the possibilities for extensions and improvements of our approach are numerous and well beyond a scope of a single paper. However, even in its current form, we provide a significant advance by combining machine learning with fundamental biodiversity scaling theory in a user-friendly concept.

Alternatives. There are indeed alternative techniques that have been used to standardize, or integrate, heterogeneous biodiversity data. The first is individual-based or sample-based *rarefaction* (Gotelli & Colwell, 2001), which has the advantage that it can be interpreted mechanistically; however, it does not account for sample shape and area, it cannot predict species composition, nor interpolate to regions that lack data. The other approach—*integrated species distribution modlling* (ISDM)—has recently been reviewed by Miller *et al.* (2019) and Isaac *et al.* (2019). It assumes that there is a continuous and scale-free layer of point pattern intensity, which then translates to likelihood of different data types. ISDM can be used to make predictions at any desired grain, and they can be stacked across multiple species, or to make predictions of biodiversity change, although ISDMs have not yet been tested for the latter. The downsides of ISDM are strict requirements on data format, the need of having the model structure a priori, nontrivial implementation, and very high computational cost. In contrast, our approach works across any range of spatial grains and it is simpler and easier to use.

Conclusion. To conclude, we present a user-friendly framework that can account for multiple biases and data deficiencies simultaneously. It can use these deficient data to estimate: (1)

biodiversity change across spatial and temporal grains, (2) average biodiversity change, (3) change in individual species distributions. Additionally, (4) it can be used to create geographic maps of biodiversity change, loss, and gain, (5) and since it can predict changes in species composition, it can also be used to map temporal species turnover. As such, our approach can be a starting point for integrative and comprehensive analyses of multiple facets of biodiversity change. It will be useful particularly in areas where standardized and temporally replicated data are rare, particularly in the tropics.

Acknowledgements

P. K. was supported by REES (Research Excellence in Environmental Sciences) grant from Faculty of Environmental Sciences, Czech University of Life Sciences in Prague. J. M. C. gratefully acknowledges the support of iDiv funded by the German Research Foundation (DFG—FZT 118, 202548816).

References

Alroy, J. (2015). Current extinction rates of reptiles and amphibians. *Proceedings of the National Academy of Sciences*, *112*(42), 13003–13008.

Azaele, S., Cornell, S. J., & Kunin, W. E. (2012). Downscaling species occupancy from coarse spatial scales. *Ecological Applications*, *in press*.

Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial point patterns: Methodology and applications with R*. CRC Press, Taylor & Francis Group.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57. https://doi.org/10.1038/nature09678

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., Moyes, F., Magurran, A., McGill, B., Myers-Smith, I. H., Winter, M., Bjorkman, A. D., Bowler, D. E., Byrnes, J. E. K., Gonzalez, A., Hines, J., Isbell, F., Jones, H. P., Navarro, L. M., … Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. *Science*, *366*(6463), 339–345. https://doi.org/10.1126/science.aaw1620

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (Eds.). (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Brummitt, R. K. (2001). *World geographical scheme for recording plant distributions, Edition 2. Biodiversity Information Standards (TDWG)*. http://www.tdwg.org/standards/109

Chase, J. M., McGill, B. J., Thompson, P. L., Antão, L. H., Bates, A. E., Blowes, S. A., Dornelas, M., Gonzalez, A., Magurran, A. E., Supp, S. R., Winter, M., Bjorkman, A. D., Bruelheide, H., Byrnes, J. E. K., Cabral, J. S., Elahi, R., Gomez, C., Guzman, H. M., Isbell, F., … O'Connor, M. (2019). Species richness change across spatial scales. *Oikos*, *128*(8), 1079–1091. https://doi.org/10.1111/oik.05968

Dormann, C. F., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmak-

er, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., & Magurran, A. E. (2014). Assemblage time series reveal biodiversity change but not systematic loss. *Science*, *344*(6181), 296–299. https://doi.org/10.1126/science.1248484

Englund, G., & Hambäck, P. A. (2007). Scale dependence of immigration rates: Models, metrics and data. *Journal of Animal Ecology*, *76*(1), 30–35. https://doi.org/10.1111/j.1365-2656.2006.01174.x

Essl, F., Moser, D., Dirnböck, T., Dullinger, S., Milasowszky, N., Winter, M., & Rabitsch, W. (2013). Native, alien, endemic, threatened, and extinct species diversity in European countries. *Biological Conservation*, *164*, 90–97. https://doi.org/10.1016/j.biocon.2013.04.005

Faurby, S., Davis, M., Pedersen, R. Ø., Schowanek, S. D., Antonelli, A., & Svenning, J. (2018). PHYLACINE 1.2: The phylogenetic atlas of mammal macroecology. *Ecology*, *99*(11), 2626–2626. https://doi.org/10.1002/ecy.2443

Faurby, S., & Svenning, J.-C. (2015). Historic and prehistoric human-driven extinctions have reshaped global mammal diversity patterns. *Diversity and Distributions*, *21*(10), 1155–1166. https://doi.org/10.1111/ddi.12369

Foote, M. (1994). Temporal variation in extinction risk and temporal scaling of extinction metrics. *Paleobiology*, *20*(4), 424–444.

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*(4), 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction, second edition* (2nd ed. 2009. Corr. 7th printing 2013 edition). Springer.

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2019). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Janitza, S., Celik, E., & Boulesteix, A.-L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*(4), 885–915. https://doi.org/10.1007/s11634-016-0276-4

Jung, M., Rowhani, P., & Scharlemann, J. P. W. (2019). Impacts of past abrupt land change on local biodiversity globally. *Nature Communications*, *10*(1), 5474. https://doi.org/10.1038/s41467-019-13452-3

Keil, P., Biesmeijer, J. C., Barendregt, A., Reemer, M., & Kunin, W. E. (2011). Biodiversity change is scale-dependent: An example from Dutch and UK hoverflies (Diptera, Syrphidae). *Ecography*, *34*, 392–401. https://doi.org/10.1111/j.1600-0587.2010.06554.x

Keil, P., & Chase, J. M. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390–399. https://doi.org/10.1038/s41559-019-0799-0

Keil, P., Pereira, H. M., Cabral, J. S., Chase, J. M., May, F., Martins, I. S., & Winter, M. (2018). Spatial scaling of extinction rates: Theory and data reveal nonlinearity and a major upscaling and downscaling challenge. *Global Ecology and Biogeography*, *27*(1), 2–13. https://doi.org/10.1111/geb.12669

Kunin, W. E. (1997). Sample shape, spatial scale and species counts: Implications for reserve design. *Biological Conservation*, *82*(3), 369–377. https://doi.org/10.1016/S0006-3207(97)00042-6

Kunin, W. E. (1998). Extrapolating species abundance across spatial scales. *Science*, *281*(5382), 1513–1515. https://doi.org/10.1126/science.281.5382.1513

Lomolino, M. V., Riddle, B. R., Whittaker, R. J., & Brown, J. H. (2010). *Biogeography* (4th ed.). Sinauer Associates.

Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, *90*(4). https://doi.org/10.1002/ecm.1422

McGill, B. J. (2010). Towards a unification of unified theories of biodiversity. *Ecology Letters*, *13*(5), 627–642. https://doi.org/10.1111/j.1461-0248.2010.01449.x

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*, 8221.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Mora, C., Tittensor, D. P., & Myers, R. A. (2008). The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1631), 149–155.

Powell, K. I., Chase, J. M., & Knight, T. M. (2013). Invasive plants have scale-dependent effects on diversity by altering species-area relationships. *Science*, *339*(6117), 316–318.

Schiltz, F., Masci, C., Agasisti, T., & Horn, D. (2018). Using regression tree ensembles to model interaction effects: A graphical approach. *Applied Economics*, *50*(58), 6341–6354. https://doi.org/10.1080/00036846.2018.1489520

Storch, D. (2016). The theory of the nested species–area relationship: Geometric foundations of biodiversity scaling. *Journal of Vegetation Science*, *27*(5), 880–891. https://doi.org/10.1111/jvs.12428

van Kleunen, M., Dawson, W., Essl, F., Pergl, J., Winter, M., Weber, E., Kreft, H., Weigelt, P., Kartesz, J., Nishino, M., Antonova, L. A., Barcelona, J. F., Cabezas, F. J., Cárdenas, D., Cárdenas-Toro, J., Castaño, N., Chacón, E., Chatelain, C., Ebel, A. L., … Pyšek, P. (2015). Global exchange and accumulation of non-native plants. *Nature*, *525*(7567), 100–103. https://doi.org/10.1038/nature14910

van Kleunen, M., Pyšek, P., Dawson, W., Essl, F., Kreft, H., Pergl, J., Weigelt, P., Stein, A., Dullinger, S., König, C., Lenzner, B., Maurel, N., Moser, D., Seebens, H., Kartesz, J., Nishino, M., Aleksanyan, A., Ansong, M., Antonova, L. A., … Winter, M. (2019). The Global Naturalized Alien Flora (GloNAF) database. *Ecology*, *100*(1), e02542. https://doi.org/10.1002/ecy.2542

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., Frenne, P. D., Verheyen, K., & Wipf, S. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings of the National Academy of Sciences*, *110*(48), 19456–19459. https://doi.org/10.1073/pnas.1312779110

Viana, D. S., Keil, P., & Jeliazkov, A. (2019). Disentangling spatial and environmental effects: Flexible methods for community ecology and macroecology. *EcoEvoRxiv*. EcoEvoRxiv. https://doi.org/10.1101/871251

Wagner, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, *15*(1), 1625–1651.

Whittaker, R. J., Nogués-Bravo, D., & Araújo, M. B. (2006). Geographical gradients of species richness: A test of the water-energy conjecture of Hawkins et al. (2003) using European data for five taxa. *Global Ecology and Biogeography*, *16*(1), 76–89. https://doi.org/10.1111/j.1466-822X.2006.00268.x

Supplementary Material

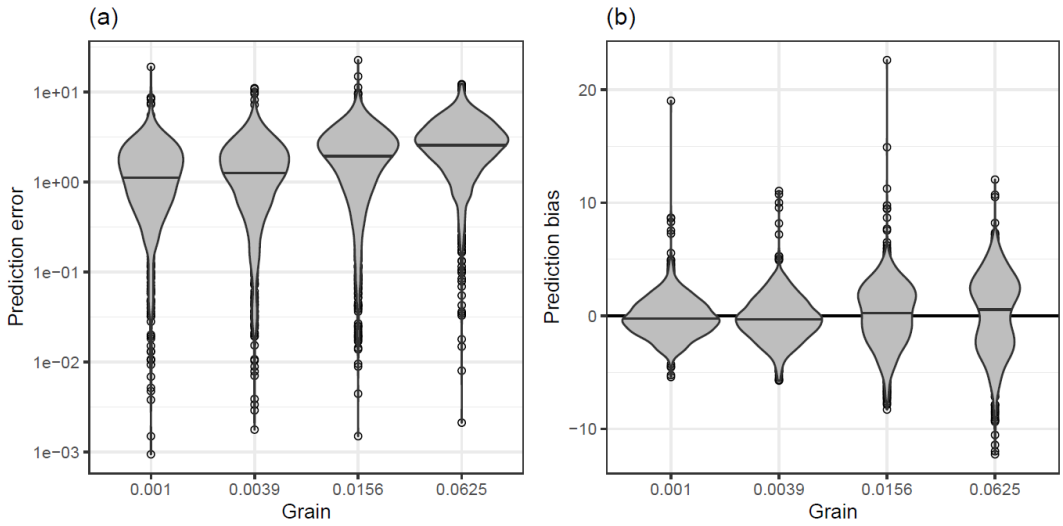

Figure S1 *Prediction error (a) and bias (b) as a function of spatial grain at which predictions were made. The errors and biases were obtained from applying boosted regression trees, and the SAR-based eq. 6 (main text) on 4,000 simulated datasets.*

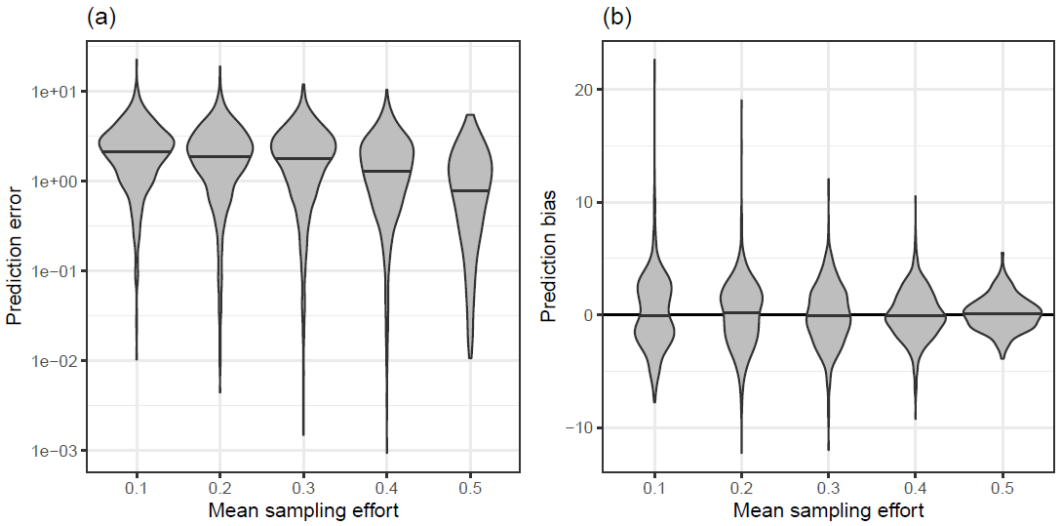

Figure S2 *Prediction error (a) and bias (b) as a function of mean sampling effort. Mean sampling effort is the proportion of polygons sampled in both time periods. The errors and biases were obtained from applying boosted regression trees, and the SAR-based eq. 6 (main text) on 4,000 simulated datasets.*

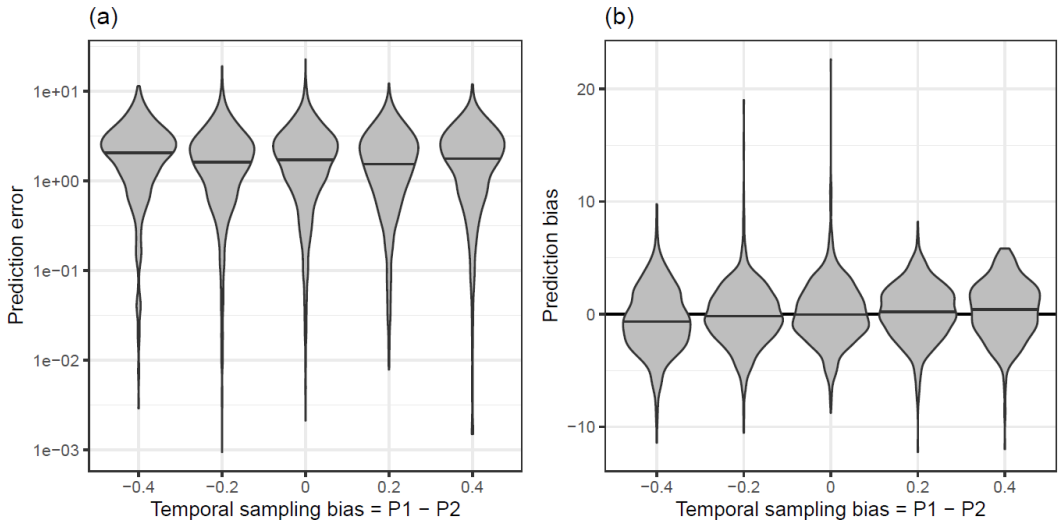

Figure S3 *Prediction error (a) and bias (b) as a function of temporal sampling bias, which is the difference between the proportions P1 and P2 of polygons sampled in time 1 and 2 respectively. The errors and biases were obtained from applying boosted regression trees, and the SAR-based eq. 6 (main text) on 4,000 simulated datasets.*