ECOGRAPHY

*Research*

# Crowd-sourced plant occurrence data provide a reliable description of macroecological gradients


**Miguel D. Mahecha, Michael Rzanny, Guido Kraemer, Patrick Mäder, Marco Seeland and Jana Wäldchen**

*M. D. Mahecha (https://orcid.org/0000-0003-3031-613X) ✉ (miguel.mahecha@uni-leipzig.de) and G. Kraemer (https://orcid.org/0000-0003-4865-5041), Remote Sensing Centre for Earth System Research, Univ. Leipzig, Leipzig, Germany. MDM also at: German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Leipzig, Germany. – M. Rzanny and J. Wäldchen, Max Planck Inst. for Biogeochemistry, Jena, Germany. – P. Mäder and M. Seeland, Dept of Computer Science and Automation, Ilmenau Univ. of Technology, Germany.*

**Ecography**
**44: 1131–1142, 2021**
doi: 10.1111/ecog.05492

Subject Editor: Joaquin Hortal
Editor-in-Chief: Miguel Araújo
Accepted 8 April 2021

Deep learning algorithms classify plant species with high accuracy, and smartphone applications leverage this technology to enable users to identify plant species in the field. The question we address here is whether such crowd-sourced data contain substantial macroecological information. In particular, we aim to understand if we can detect known environmental gradients shaping plant co-occurrences. In this study we analysed 1 million data points collected through the use of the mobile app Flora Incognita between 2018 and 2019 in Germany and compared them with Florkart, containing plant occurrence data collected by more than 5000 floristic experts over a 70-year period. The direct comparison of the two data sets reveals that the crowd-sourced data particularly undersample areas of low population density. However, using nonlinear dimensionality reduction we were able to uncover macroecological patterns in both data sets that correspond well to each other. Mean annual temperature, temperature seasonality and wind dynamics as well as soil water content and soil texture represent the most important gradients shaping species composition in both data collections. Our analysis describes one way of how automated species identification could soon enable near real-time monitoring of macroecological patterns and their changes, but also discusses biases that must be carefully considered before crowd-sourced biodiversity data can effectively guide conservation measures.

Keywords: automated species identification, canonical correlation analysis, citizen science, floristic survey, macroecological patterns, nonlinear dimensionality reduction

## Introduction

Climate change, habitat losses and intensified land-use dynamics threaten current levels of biodiversity across the planet (Blowes et al. 2019, Brondizio et al. 2019). Accurately predicting the effects of these global transformations on species occurrences, communities and ultimately on ecosystem functioning, has made substantial progress over the last decades, but uncertainties remain high as long as reference data are scarce (Urban et al. 2016). It is therefore important to find new ways to quickly and reliably monitor species (co-)occurrences in space and time.

© 2021 The Authors. Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.



NORDIC SOCIETY OIKOS

www.ecography.org

Automated species identification is one promising avenue that has been discussed since the potential of machine learning for ecological applications has become evident (Gaston and O'Neill 2004). With the advent of deep learning methods (Goodfellow et al. 2016) automated species identification reaches levels of accuracy comparable to human experts (Bonnet et al. 2018, Wäldchen and Mäder 2018a, Jones 2020, Villon et al. 2020). Today, multiple smartphone apps leverage such algorithms and enable users to identify e.g. plants, insects or birds directly in the field (Kumar et al. 2012, Affouard et al. 2017, Van Horn et al. 2018, Wäldchen and Mäder 2018a, Jones 2020). The voluntarily shared ancillary information on time and location could soon turn such mobile observations into an invaluable resource for different monitoring tasks (Bonnet et al. 2020). Consequently, the questions discussed in the existing literature on automated species identification mainly revolve around the accuracy of the competing algorithmic approaches for automated species identification (Nguyen et al. 2018, Wäldchen and Mäder 2018b, Jones 2020, Villon et al. 2020).

Despite the popularity of the corresponding apps for automatic species identification, little is known about the joint potential of such individual observations to reveal biodiversity gradients and patterns of plant co-occurrences. If it could be shown, however, that data collected with the help of mobile apps can be used identify specific macroecological patterns, a wide range of new perspectives for the study of large-scale biodiversity dynamics would emerge. The aim of this study is to take a step in this direction. We want to understand whether an opportunistic crowd-sourced data collection of two years, derived from a smartphone application for automatic plant identification, encodes macroecological patterns of plant co-occurrences similar to those that can be obtained from data collected over decades of intensive mapping of plant species in the same geographical area. We used the records of the plant identification app Flora Incognita, an app developed to automatically identify all vascular plant species in Germany, which is increasingly used throughout Europe (<https://floraincognita.com/>, Wäldchen et al. 2018, Rzanny et al. 2019, Mäder et al. accepted) and compared them to the records of the German sampling program Florkart, which is a long-term cumulative mapping project that turned into a national sampling effort administered by the German Federal Agency for Nature Conservation (Netzwerk Phytodiversität Deutschland and Bundesamt für Naturschutz 2013).

The study is structured as follows: firstly, we compare the two data collections at the grid level and investigate potential biases. Specifically, we investigated whether human population density, a known bias in crowd-sourced data, would have an imprint on the data quality. Secondly, we focus on comparing the macroecological co-occurrence patterns in the observed species compositions extracted with non-linear dimensionality reduction. The rationale is that both data sets can be regarded as independent samples from the same ecological gradients, hence we would expect that the extracted patterns are directly related. Thirdly, we explore the joint low-dimensional ecological gradients in relation to climatic and soil properties in order to understand whether the recovered co-occurrence patterns are indeed related to expected environmental drivers. Finally, we discuss data limitations and our findings in the light of the urgently needed perspectives for rapid assessments of the state of biodiversity and macroecological patterns beyond the aspects investigated here.

## Material and methods

### Data

#### *Flora Incognita*

Flora Incognita (<https://floraincognita.com/>, Mäder et al. 2021) is a freely available mobile app allowing the automated identification of wild flowering plants and originally developed for the German flora. In January 2020 the app was already able to identify 4848 vascular plant species, covering the central European flora and beyond. Depending on the difficulty of identification, the app analyses one or more smartphone photograph(s) from predefined perspectives. Images of the whole plant or plant organs such as flowers, leaves or fruits are incrementally transferred to the Flora Incognita server until the plant can be identified at the species level and the result is transferred back to the user's device. The interactive classifier utilises a task-specific cascade of convolutional neural networks (CNN), a default choice for analysing high-dimensional and spatially correlated input data such as images (LeCun et al. 1995, Wäldchen and Mäder 2018a). Taxonomy for species occurring in Germany is based on GermanSL (Jansen and Dengler 2008) with some critical genera (e.g. *Hieracium*, *Rubus*, *Sorbus*, *Taraxacum*) not fully resolved to the species level. Depending one the score of the result, the classifier requires one to three images of different perspectives or organs. The classifier is pre-trained on the ImageNet data set and fine-tuned on the growing Flora Incognita data set; today the training data set contains approximately two million training images. Identifications suggested by the classifier need to be confirmed by the user in order to create a record. Upon confirming the identification, the result is stored in the database including images, date and geographical location – if provided voluntarily by the user. For the comparison of the Flora Incognita to the national inventory (Florkart, see below), we used only confirmed records including location data and transferred all data points into presence–absence records referring to the same geographical grid as used by Florkart. For this study we used all data points with geolocation in Germany collected in 2018 and 2019 summing up to 961 116 records.

#### *Florkart*

As reference data we used Florkart, the inventory of vascular plant ocurrences in Germany. The database contains multiple mapping campaigns involving thousands of voluntary surveyors as well as literature reviews that were gathered since the middle of last century. The data are freely accessible via the information system FloraWeb (<www.floraweb.de/>)

run by the Federal Agency for Nature Conservation on behalf of the German Network for Phytodiversity (NetPhyD; Netzwerk Phytodiversität Deutschland and Bundesamt für Naturschutz 2013). The presence of a species is recorded on the basis of grid tiles, originally representing ordnance survey maps at a scale of 1:25 000. Each tile covers a section of 10′ longitude × 6′ latitude, corresponding to a surface area of approximately 118 km$^2$ in the north to 140 km$^2$ in the south of Germany. Each tile carries the binary information whether a species appears in it or not. Neither exact spatial coordinates of individual records nor frequencies are known. Today, Florkart comprises approx. 30 000 000 individual records. Both, the crowd-sourced Flora Incognita data and long-term inventory data of Florkart were preprocessed to contain only species names and grid cells shared by both data sets. Intersecting the 2761 Florkart dimensions and Flora Incognita results in 2598 species that potentially occur in 3003 spatial grid cells.

#### *Ancillary data*

Unlike Florkart, we had to assume that the number of Flora Incognita records would strongly depend on the number of device carriers i.e. population density at a certain location. In order to understand the potential bias due to population density in the direct comparison of Flora Incognita and Florkart we used population data from the German National Census 2011, <https://www.zensus2011.de/EN/> as reference. These data were provided by the Federal Statistical Office.

To explain macroecological patterns we used the bioclimatological features and standard variables from WorldClim (<www.worldclim.org/>, Fick and Hijmans 2017, highest resolution 30 s, 103 variables in total), as well as soil physical and chemical properties derived from Soil Grids (<https://soilgrids.org/>, Hengl et al. 2014, provided at 250 m spatial resolution, 127 variables). All environmental variables were aggregated to the grid cells predefined by Florkart, where we averaged values where appropriate, but took the grid cell minimum or maximum where the variables required that (e.g. maximum/minimum temperatures of warmest/coldest month). We then subjected the variables belonging to WorldClim and SoilGrids to a principal component analysis (PCA; cf. Supporting information) in order to obtain a reduced set of interpretable predictors for both the climate and edaphic domain. We further worked with the four leading principal components for WorldClim and SoilGrids as the differences in explained variance in the scree plot for subsequent axes was marginal (i.e. after the 'elbow'). This heuristic approach leads to two sets of PCs that account for 89.8% (WorldClim) and 90.6% (Soil Grids) of the variance respectively, leading to well clearly identifiable components that were used in the subsequent analyses.

### Dimensionality reduction

The preprocessed species occurrence data are two data arrays $\mathbf{X}_{FK}$, $\mathbf{X}_{FI} \in \{0,1\}^{n \times p}$, where the indices FK and FI indicate either Florkart or Flora Icognita respectively, and $\mathbf{X}$ contains binary information on the presences/absence of $p = 2598$ species at $n = 3003$ spatial grid cells. The general idea is that Flora Incognita and Florkart data sets are both (noisy) samples from the same unknown environmental gradients (the so-called underlying manifold). In other words, we assume that the $p$ species are highly redundant meaning that the presence–absence pattern of one species might probably be very similar, yet not identical, to multiple other species. This argument leads to the expectation that we can find some $q_{FK}$ and $q_{FI}$-dimensional representation of these data, i.e. $\mathbf{Y}_{FK} \in \mathbb{R}^{n \times q_{FK}}$ and $\mathbf{Y}_{FI} \in \mathbb{R}^{n \times q_{FI}}$, that retain significant proportions of variance. These orthogonal, non-redundant, dimensions shall represent the empirical macroecological patterns we want to compare. We note that $q_{FK}$ and $q_{FI}$ can be different, depending on the performance of the dimensionality reduction for each of the data sets. To retrieve these underlying dimensions, we performed nonlinear dimensionality reduction separately for each data set via isometric feature mapping (Isomap; Tenenbaum et al. 2000). The method is essentially a classical multidimensional scaling (CMDS; Legendre and Legendre 2012), but instead of aiming for preserving a matrix of linear ecological distances among the n locations, it tries to preserve the geodesic ecological distances (Mahecha et al. 2007). To estimate the geodesic distances we initially compute a distance matrix, $\mathbf{D} \in \mathbb{R}^{n \times n}$, based on the Jaccard metric

$$d(\mathbf{a},\mathbf{b}) = 1 - \frac{|\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|} \in [0;1]$$

where $\mathbf{a},\mathbf{b} \in \{0,1\}^p$ are vectors of presence absence data (Jaccard 1901). Using a $k$-nearest neighbour ($k$-NN) graph, we can compute the shortest path among data samples. To find an optimal $k$-value we iterated from its minimal plausible value to theoretical maximum (where Isomap becomes a standard CMDS) and explored how much variances can be explained by the leading dimensions (results of this sensitivity analysis are provided in the Supporting information). This pre-analysis shows that best data-compression results are obtained for relatively low $k$-values, pointing at a highly nonlinear underlying data space. For all embeddings with $k < 40$ we then performed a canonical correlation analysis (CCorA) between the embeddings of Flora Incognita and Florkart and identified the $k = 16$ as the one where canonical correlations where highest (Supporting information), while the compression of the individual data sets is optimal (Supporting information). All subsequent analysis are based on these two embeddings. The implementation followed the concept outlined by Kraemer et al. (2018). To quantify the autocorrelation in the Isomap components we used Moran's $I$ at lag one, based the queen adjacency neighbourhood definition.

### Canonical correlation analysis, CCorA

The application of dimensionality reduction on the two data sets of interest leads to the matrices $\mathbf{X}_{FK}$ and $\mathbf{Y}_{FI}$, i.e. $n$ samples in each of the $q_{FK}$ or $q_{FI}$ dimensions. More formally, we compared two data sets that were assumed to be samples from the same underlying manifold. And, if this were the case, we could expect the recovered dimensions to be almost identical

and linearly related even if the dimensions themselves have been extracted by nonlinear dimensionality reduction. The whole idea behind Isomap is that it can flatten the nonlinearily encoded manifold (Tenenbaum et al. 2000). We therefore sought the canonical correlation patterns that would maximize the relation of both recovered sets of dimensions. Canonical correlation analysis (CCorA) seeks linear combinations of the two input data sets (here the sets of Isomap dimensions $\mathbf{X}_{\mathrm{FK}}$ and $\mathbf{X}_{\mathrm{FI}}$) such that their common correlation is maximized (Legendre and Legendre 2012). In other words, for the first canonical variate we seek $\mathbf{z}_{\mathrm{FK}}^{(1)} = \mathbf{Y}_{\mathrm{FK}}\mathbf{v}_{\mathrm{FK}}^{(1)}$, and $\mathbf{z}_{\mathrm{FI}}^{(1)} = \mathbf{Y}_{\mathrm{FI}}\mathbf{v}_{\mathrm{FI}}^{(1)}$, with $\mathbf{z}_{\mathrm{FK}}^{(1)}, \mathbf{z}_{\mathrm{FI}}^{(1)} \in \mathbb{R}^n$, $\mathbf{v}_{\mathrm{FK}}^{(1)} \in \mathbb{R}^{q_{\mathrm{FK}}}$, $\mathbf{v}_{\mathrm{FI}}^{(1)} \in \mathbb{R}^{q_{\mathrm{FI}}}$, such that they maximize their correlation $\rho_1 = \mathrm{corr}\left(\mathbf{z}_{\mathrm{FK}}^{(1)}, \mathbf{z}_{\mathrm{FI}}^{(1)}\right)$. The emerging variables $\mathbf{z}_{\mathrm{FK}}^{(1)}$ and $\mathbf{z}_{\mathrm{FI}}^{(1)}$ are said to be the first canonical variates, and all subsequent canonical variates, $\mathbf{z}_{\mathrm{FI}}^{(i)}$ and $\mathbf{z}_{\mathrm{FK}}^{(i)}$, $1 \leq i \leq \min(q_{\mathrm{FK}}, q_{\mathrm{FI}})$ are orthogonal to the previous. The solution to the problem can be achieved via singular value decomposition of the cross-correlation matrices. Bartlett's $\chi^2$-test reveals the significance of the canonical correlations.

### Predicting spatial patterns

In order to predict the leading Isomap dimensions of Florkart and Flora Incognita, as well as the canonical variates, we used a random forest approach (Breiman 2001) that has been widely used i.e. for ecological applications (Pompe et al. 2008, Bodesheim et al. 2018). This flexible prediction approach can cope with correlations among the predictors and nonlinearities in the relation of predictors to target variables. Cross validation is needed to identify random forest regression models and estimate variable importances. However, because our data are strongly autocorrelated in space a random selection of data points for cross validation would lead to overfitted models (Brenning 2012). One approach to control for this effect is to leave spatially contiguous blocks of samples out in each cross-validation step. It has been recently suggested not to select blocks based either on geographical distances, but to clustering spatial points by their environmental conditions to obtain 'environmentally separated folds' (Valavi et al. 2018). Here we combined the idea of clustering by geographical and environmental conditions. This was possible using the affinity propagation algorithm (Frey and Dueck 2007) for data clustering, which works using asymmetric similarity matrices. We constructed a similarity matrix with the lower triangle contained environmental similarities of points estimated in the space of the four leading Isomap dimensions, and the upper counterpart reflects the geographical proximity (inverse to the geographical distance). The geographical proximities penalises points that are environmentally similar, but geographically distant to each other in the search of the clusters. The advantage of affinity propagation is that the number of clusters emerges from the data and does not have to be defined a priori (result of the clustering are shown in the Supporting information). The cross validation using these clusters as spatial labels, together with variable selection and estimating variable importance was performed with the 'caret applications for spatio–temporal models' (CAST R package) approach following Meyer et al. (2018). Because this study does not contain time information, it essentially reduces to a spatial cross-validation sensu (Brenning 2012).

## Results

### Direct comparison of Flora Incognita and Florkart

A naïve comparison of both data sets based on Jaccard dissimilarity at the grid-cell level revealed strongest differences in the least densely populated regions of Germany (Fig. 1a) – a bias we expected as a result of differences in the frequency of app usage. Jaccard dissimilarities between Florkart and Flora Incognita ranged between 0.55 and 1 (median of 0.87; Fig. 1a), which means that at best 45% of species occurrences were observed in both data sets across some grid cells, but some cells shared essentially no common species observation. The Jaccard dissimilarities showed a clear pattern of spatial clustering with lower values (i.e. better correspondence) around Berlin, Hamburg, Bremen and other densely urbanized regions. We observed maximum dissimilarities between the two collections in rural areas. Fig. 1d shows the Jaccard dissimilarities related to population counts. This visualization suggests that for each level of population count we can estimate a maximum dissimilarity between Florkart and Flora Incognita, and that this dissimilarity decreases with increasing population density. In other words, the correspondence of species composition among the two data sets is expected to increase with population density. The quality of Flora Incognita coverage seems to be primarily dependent on the probability that a smartphone user is at a particular location. However, we also found deviations from this pattern. We labelled the grid cells that showed a much lower Jaccard dissimilarity than the population count led us to expect. These places included, for example, Jena and Ilmenau, where the app was developed, but also well-known destinations for tourism (e.g. Zugspitze, Germany's highest mountain; Amrum, an island in the North Sea). These findings suggest that domestic ecotourism is beneficial for collecting mobile-assisted citizen science data.

The Jaccard dissimilarities between Florkart and Flora Incognita are a symmetric measure of data mismatch (Fig. 1a). Hence, the observed levels of dissimilarity stem from a systematic undersampling from either of the two data sets. Figure 1b shows the Florkart exceedance, which is the ratio of species recorded in Florkart to those contained in both data sets, and Fig. 1c shows the corresponding Flora Incognita exceedance levels. This analysis indicated that the sampling biases were entirely owing to Flora Incognita (median of 7.04; Fig. 1b). Again, this asymmetric bias related well to the population counts (Fig. 1e). There were very few places where Flora Incognita indicated more species occurrences than Florkart, and these were typically at the national border where Florkart has almost no records (Fig. 1c).

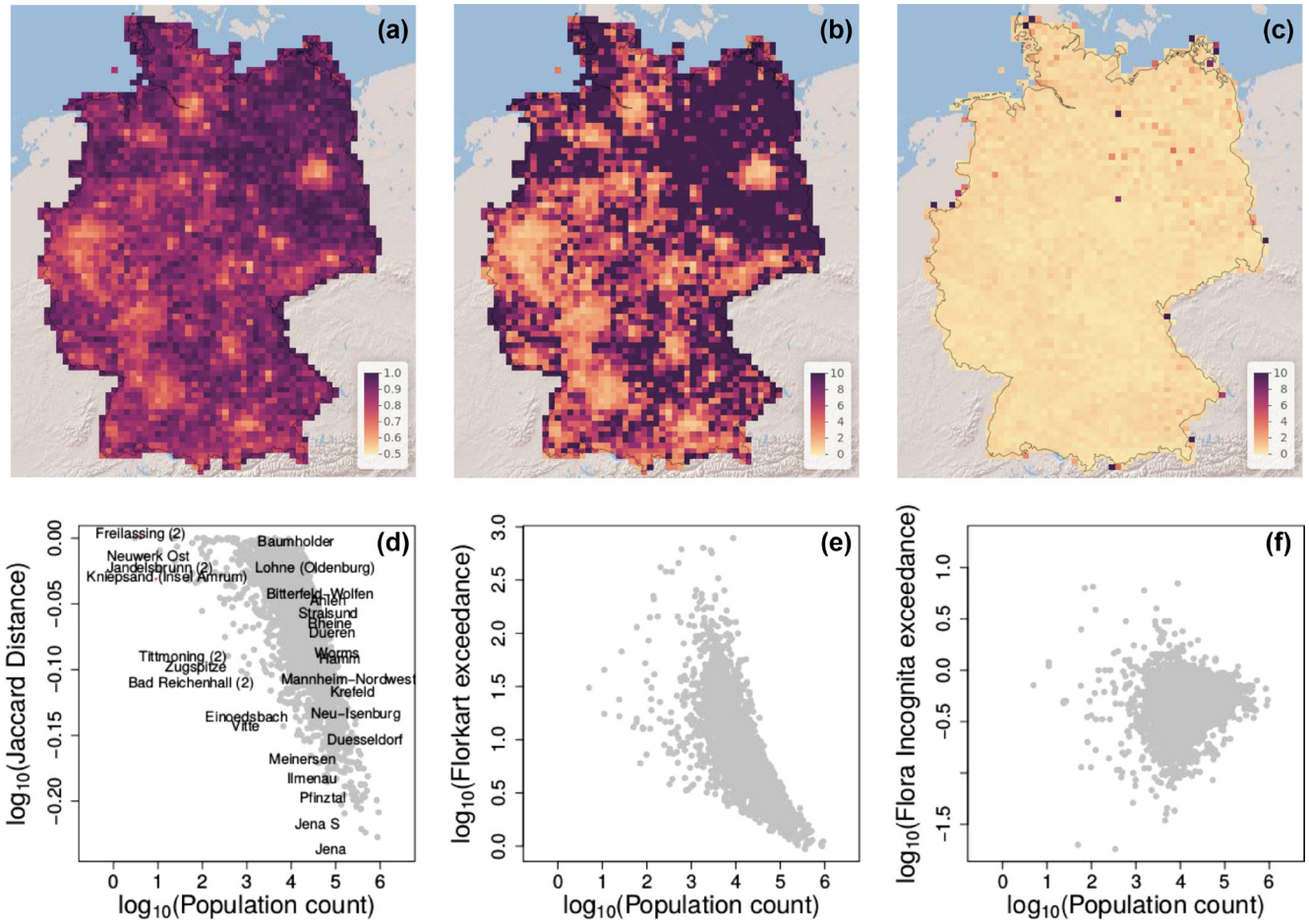


Figure 1. Spatial inconsistencies between the Flora Incognita app data and the long term national inventory Florkart. (a) Map of Jaccard-distances between Florkart and Flora Incognita occurrence data at the grid-cell level (observed minimum = 0.55). Maximal distances are found in rural areas in north-eastern-Germany. (b) Florkart exceedance level: the fraction of species in Florkart that are not found by Flora Incognita; and analogously. (c) Flora Incognita exceedance level. (d–f) Relation of the values in the maps to population counts from the national census 2011.

## Low-dimensional representations

We assumed that the $p = 2598$ species considered in this study were samples from a low-dimensional manifold, whose intrinsic dimension $q \ll p$. Fig. 2 shows the comparison of the degree to which the two data streams can be compressed. We found that Florkart could be compressed much better than Flora Incognita (residual variances in four dimensions were approx. 10% versus 40%; for a sensitivity analysis with varying parameters see the Supporting information). In the following we worked with the leading four Isomap components of both data sets, selected such that for both data set additional dimensions did not explain substantially more variance. Given that both data sets were of exactly the same extrinsic dimension and covered the same geographical and species range, this figure suggests that Flora Incognita data must be noisier (i.e. harder to compress).

Visualizing the leading components in geographic space partly supports this conjecture (Fig. 3a–d versus Fig. 3e–h). While the leading dimensions from Florkart data showed smooth geographical patterns, the analogously extracted components from Flora Incognita were less smooth. To confirm the notion of lower autocorrelation in the patterns extracted from Flora Incognita we estimated Moran's $I$ for each dimension and get the following values for Florkart $I\left(\mathbf{y}_{\mathrm{FK}}^{(1)}\right) = 0.96$, $I\left(\mathbf{y}_{\mathrm{FK}}^{(2)}\right) = 0.90$, $I\left(\mathbf{y}_{\mathrm{FK}}^{(3)}\right) = 0.75$ and $I\left(\mathbf{y}_{\mathrm{FK}}^{(4)}\right) = 0.81$, and for Flora Incognita $I\left(\mathbf{y}_{\mathrm{FK}}^{(1)}\right) = 0.57$, $I\left(\mathbf{y}_{\mathrm{FK}}^{(2)}\right) = 0.67$, $I\left(\mathbf{y}_{\mathrm{FK}}^{(3)}\right) = 0.60$ and $I\left(\mathbf{y}_{\mathrm{FK}}^{(4)}\right) = 0.53$. In light of the limited compressibility (Fig. 2), this finding is yet another hint for a high amount of noise in the crowd-sourced data.

However, we also note that the first Isomap dimension derived from Flora Incognita (Fig. 3e) apparently shares certain patterns with the undersampling biases that we related to population density (cf. Fig. 1a–b). To investigate this further, we used a machine-learning-based variable selection approach to understand to what degree 'population count' would be selected as a key predictor variable for each of the Isomap dimensions in both data sets. The results show that the dominant factor explaining the first two Flora Incognita

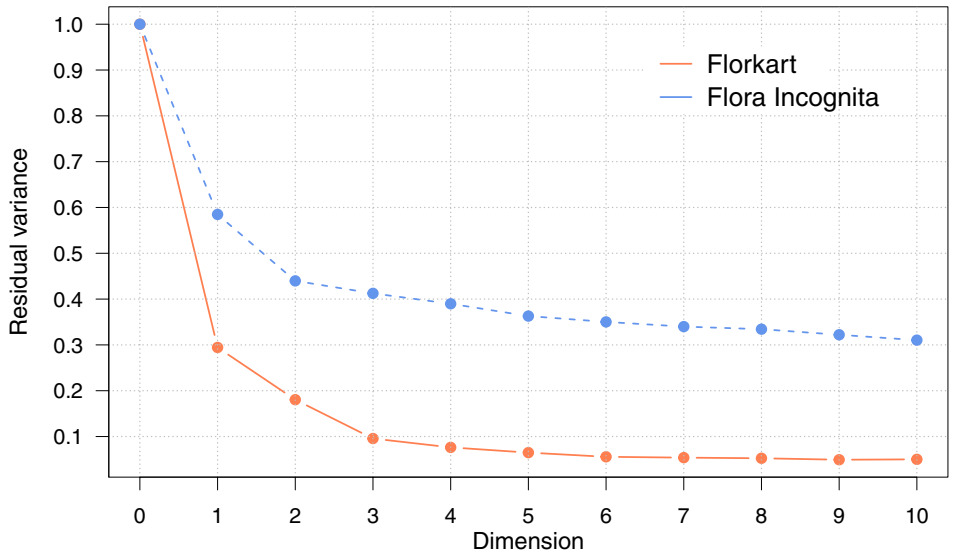


Figure 2. Effectiveness of nonlinear dimensionality reduction. Residual variances in the compression of Florkart and Flora Incognita with Isomap (here $k$-NN = 16) show that Florkart can be compressed much more strongly than Flora Incognita, suggesting that we have much reduced noise levels.

dimension is population density (Supporting information). The leading two Florkart dimensions, instead, can only be related to the two climate axes temperature and the wind/seasonality axes, while the third axis has also a strong link to population density. Given that there is no reason to expect a population density effect in Florkart one shall also consider that population density and environmental conditions are not necessarily independent from each other.

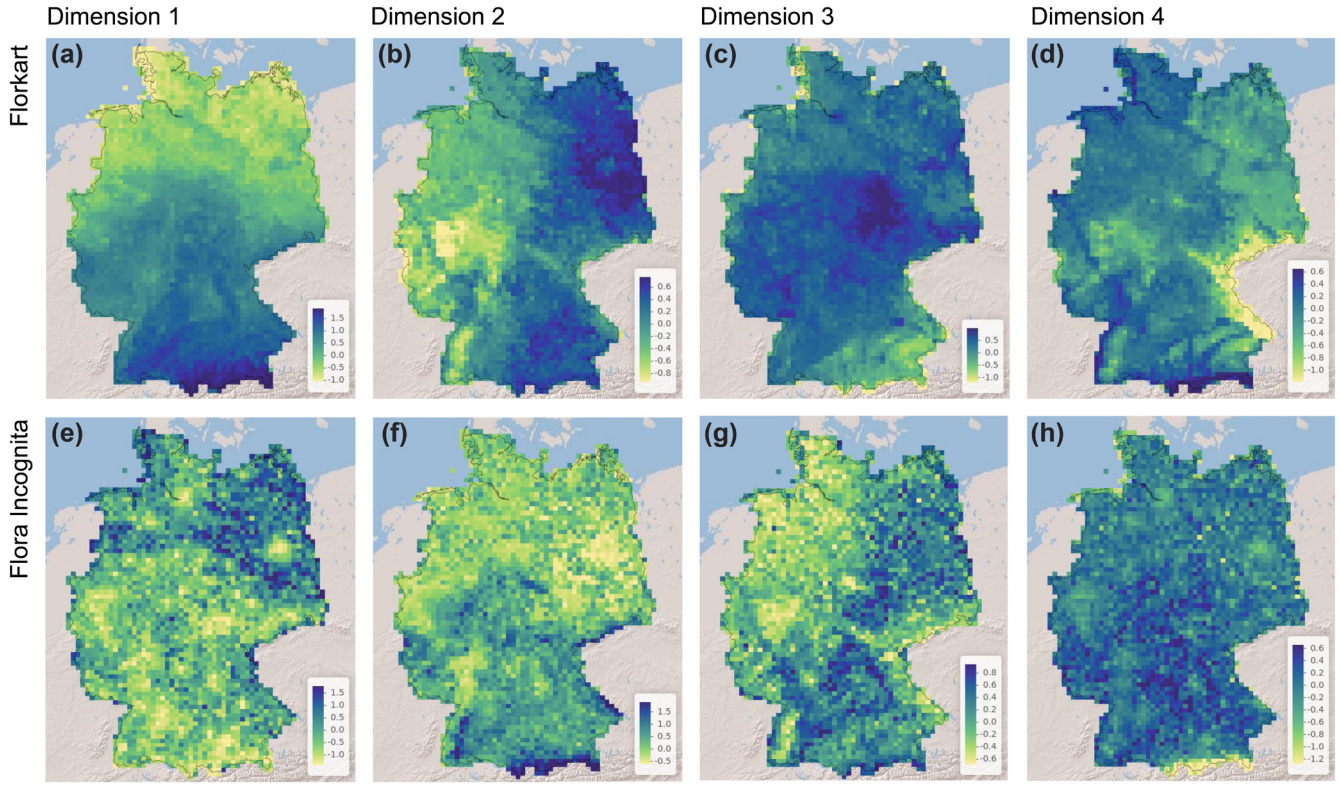


Figure 3. Leading biogeographical dimensions recovered from Florkart and Flora Incognita. (a–d) Show the leading Isomap dimensions, $\mathbf{y}_{\mathrm{FK}}^{(i)}$, and in (e–h) the analogously extracted Isomap dimensions from Flora Incognita $\mathbf{y}_{\mathrm{FI}}^{(i)}$ where $i = 1,\ldots,4$. Biogeographical gradients in Florkart are much smoother compared for Flora Incognita. The first Flora Incognita dimension in (e) is similar to the sampling bias in Fig. 1a. All others share patterns with the leading Florkart dimensions.

Interestingly, Flora Incognita's second Isomap dimension was visually comparable to the first dimension recovered from Florkart. Figure 3 led us to suspect that higher Flora Incognita dimensions corresponded to the Florkart dimension, except for the first dimension, which apparently mainly carried the signal of the spatial bias.

## Common macroecological patterns

In the best conceivable case, Flora Incognita and Florkart would have exactly the same entries for each species at each location. Hence, we next investigated the degree to which the dimensions extracted from Flora Incognita and Florkart were aligned and could be interpreted as a common set of underlying environmental conditions. Figure 4 shows the four leading canonical variables, that is, the linear combinations of Isomap dimensions from each of the two data sets that maximize mutual correlations under the constraint of sequential orthogonality. We obtained four significant canonical variates with correlation values of $\rho_1 = 0.86$, $\rho_2 = 0.65$, $\rho_3 = 0.49$ and $\rho_4 = 0.37$. These values of correspondence led us to conclude that the Flora Incognita records contain the major

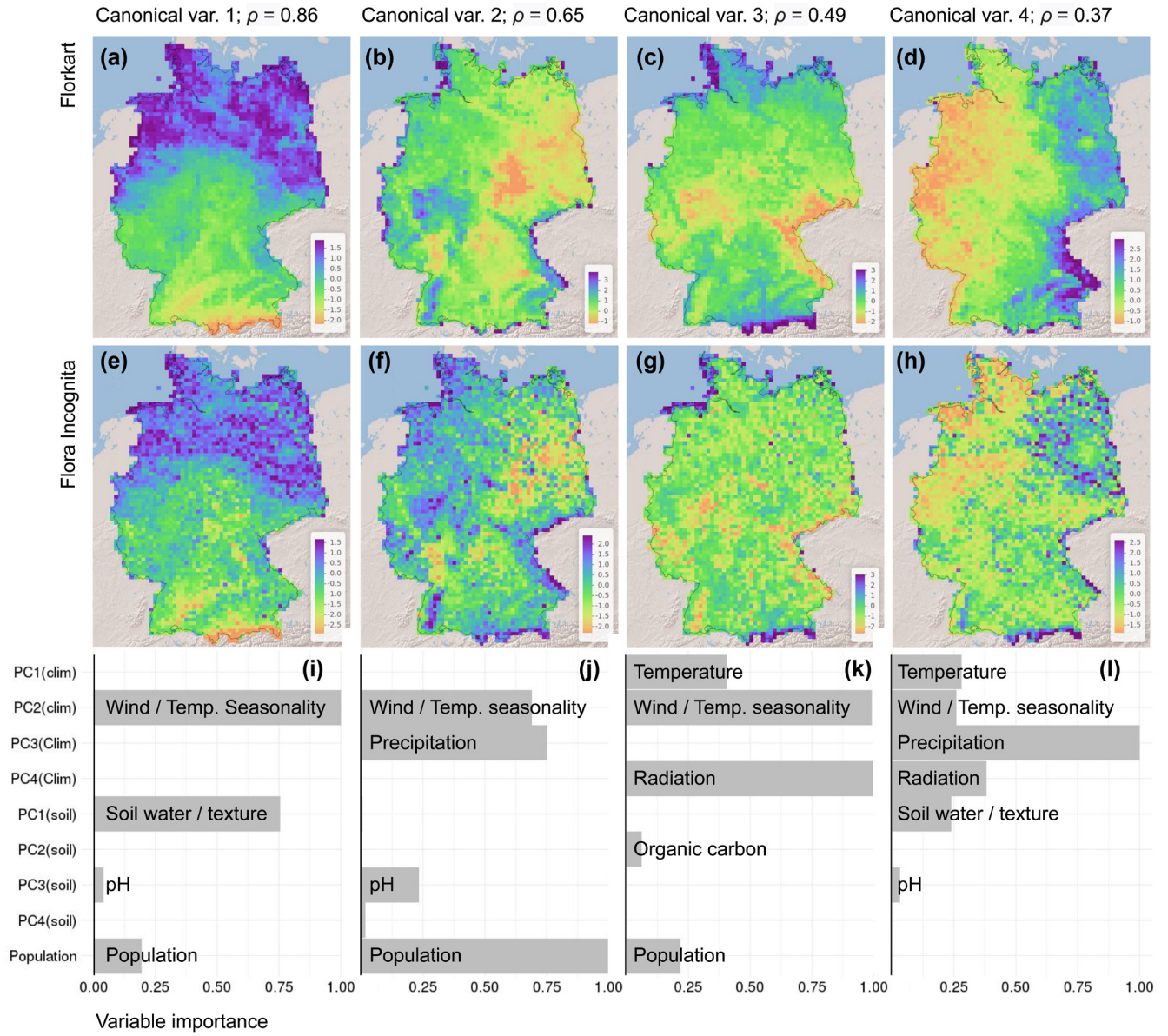


Figure 4. Joint biogeographical gradient in Flora Incognita and Florkart recovered from their leading Isomap dimensions and their relation to environmental drivers. (a–d) Significant canonical variates for Florkart $\mathbf{z}_{FK}^{(i)}$, $i = 1,\ldots,4$ and in (e–h) for Flora Incognita $\mathbf{z}_{FI}^{(i)}$, $i = 1,\ldots,4$. Both data sets share four biogeographical gradients with canonical correlations ranging from $\rho_1 = 0.86$ to $\rho_4 = 0.37$. (i–l) Predicting the joint canonical variates (e.g. $\mathbf{z}_{FK}^{(1)}$ and $\mathbf{z}_{FI}^{(1)}$) with random forests allows an environmental interpretation shown here as variable importance. Results from separate variable importance estimates for the prediction of the pairs $\mathbf{z}_{FK}^{(i)}, \mathbf{z}_{FI}^{(i)}$ leads to similar, yet not identical results (Supporting information).

biogeographical patterns of the German flora. Essentially, the canonical variate 1 was only driven by Florkart dimension 1, while in the case of Flora Incognita, the canonical variate 1 combined patterns from its first two Isomap dimensions (results in Supporting information). The second canonical variate corresponded to Isomap dimensions 3 of both Flora Incognita and Florkart.

The identified joint macroecological patterns (Fig. 4) suggest that Flora Incognita indeed captures major patterns of species composition that could be linked to climatic and edaphic conditions. Using a machine learning regression approach that accounted for nonlinear relations, which was cross-validated to consider contiguous spatial folds to avoid overfitting due to autocorrelation, we were able to interpret all four significant canonical variates. In Fig. 4i–l we show the variable importance for the joint prediction of the Florkart and Flora Incognita canonical variates. The first joint pattern is the most strongly related to the wind/temperature seasonality/temperature maximum axis, but also has a strong imprint of the soil water/texture axis. Interestingly, the first three canonical variates also reveal an imprint of population density. We suspect that in Germany, the human imprint is interrelated with natural sources of variation, as reported previously e.g. by Kühn et al. (2004), and it should always be considered a predictor beyond the analysis of data biases. The second canonical variate mainly reflects the signature of the precipitation axis, but again the wind/temperature seasonality/temperature maximum axis has an effect, along with an additional imprint of soil pH values. The third and forth axes are the most complex, containing the interactions of more climate variables. The fact that PC4(soil) is not picked up as predictor can be explained by the fact that this is yet a second water related axis, which is however more related to soils clay content and therefore nonlinearily related to PC1(soil). Results for the variables of importance in predicting the canonical variates independently are shown in the Supporting information. The models trained for the four canonical variates explained 49%, 51%, 38% and 41% of variance, respectively (in spatial cross validation).

## Discussion

This study reports two apparently contradictory findings: On the one hand we find that macroecological species composition patterns uncovered from crowd-sourced plant occurrences can reveal plausible environmental gradients (Fig. 4), on the other hand we have identified substantial spatial biases in the raw data (Fig. 1). The fact that the recovered macroecological patterns were robust to the spatial biases is an argument in favour of the chosen approach of nonlinear dimensionality reduction. However, these results must be interpreted with due caution, as the robustness of the presented results and the perspectives of the approach for comparable applications will always depend on data quality. In the following we therefore firstly discuss data quality issues, secondly the robustness of the macroecological patterns, and finally develop a perspective for transferring this approach the continental or global scale.

### Data quality of automated species identification

Today multiple apps for automated plant species identification are available (e.g. Bing, Flora Incognita, Google Lens, iPlant, PlantNet, PlantSnapb, Seek/iNaturalist; Jones 2020). One question is whether data collected by any of the other apps would haven been better, equally or less suitable for this specific study. A general remark is that Flora Incognita was launched in spring 2018 with the aim of supporting users in identifying flowering species in Germany. For the specific case of the German flora, the app's accuracy approaches > 93% for single observations (Wäldchen et al. 2018), which are values close to expert-level classification results (Rzanny et al. 2019).. This is why, from a botanical point of view, we can regard Flora Incognita as suitable for the analysed region. However, Flora Incognita has shown potential beyond this regional focus: in a recent comparison of eight apps for automated plant species identification carried out for the British flora, it was reported that Flora Incognita is on par with other apps such as Plant.id, Google Lens and Seek (Jones 2020). The study by Jones (2020) also reports, for instance, that Flora Incognita, along with Seek makes 'the fewest wrong suggestions as they tended not to give an answer where there is uncertainty' (Jones 2020). These findings suggest that other apps could have produced data of equal quality. In general, we expect that the maturity of the automated species classification algorithms to further increase in the coming years (Affouard et al. 2017, Van Horn et al. 2018, Jones 2020) through better training data in cooperation of users and botanical experts (in the case of Flora Incognita this is achieved via the companion app Flora Capture; Boho et al. 2020), by using more advanced deep learning approaches for automated plant species identification (cf. latest advances presented e.g. in Figueroa-Mata and Mata-Montero 2020, Villon et al. 2020), or by more strongly considering additional information e.g. of geographical locations as suggested for instance by Wittich et al. (2018) or using other ancillary information (Goldsmith et al. 2016, Terry et al. 2020), e.g. exploiting novel potentials of satellite remote sensing data that can encode multiple land surface properties across scales.

Improvements in classification accuracy will, however, not solve the lack of spatial sampling coverage which is probably the major obstacle to the scientific exploitation of crowd-sourced data. This problem became obvious in the direct comparison of Flora Incognita and Florkart at the grid-cell level (Fig. 1). We assume that the main effect leading to low correspondence levels between classical inventories and crowd-sourced data is the users recording behaviour. Flora Incognita is primarily used by non-experts with an interest in common or conspicuous species. Many common inconspicuous species (such as Poaceae) are barely recorded. Another effect is that even in areas with high numbers of observations, rare species tend not to be sampled by Flora Incognita, while the expert surveys included in Florkart do report such

species. These are typical biases that have been shown in citizen-science studies when non-experts are involved in data collection (Geldmann et al. 2016, Boersch-Supan et al. 2019, Johnston et al. 2020). Another problem that might appear is the 'confirmation bias'. Users of Flora Incognita need to confirm their observations which may lead to incorrectly reclassify a species. This issue is probably not so severe, as it mostly leads to errors when the classification algorithm is highly uncertain and shall actually reduce its error rates. At the same time, Flora Incognita also provides species descriptions, where plant characteristics are described and images are shown. Users can thus verify the automatic identification again, which additionally reduces the 'confirmation bias'. Still, further quality control such as plausibility checks need to be developed to increase the reliability of records and the signal to noise ratio. At the same time, it is also important to draw the user attention to specific species groups that should be recorded. This can be achieved, for example, when specific citizen science projects are carried out with Flora Incognita or comparable apps.

## On retrieving empirical macroecological patterns

Related to the question of data quality issues during crowdsourcing is need for robust reference data. In this study, we relied on Florkart which today integrates observations collected over 70 years, but suffers from some caveats as well. For instance, Mahecha and Schmidtlein (2008) reported sampling biases in Florkart due to inaccurate naming conventions in particular in times where Germany was politically divided. Recently, Eichenberg et al. (2020) reported that, with the exception of neophytes, many species groups have declined in Germany over the past decades. Their study is mainly based on Florkart data and reports a negative trends in species richness of $-0.19$ [% year$^{-1}$] after correcting for multiple potential taxonomic and sampling issues.

While we can assume that the spatial biases in Florkart are marginal to the present study (Mahecha and Schmidtlein 2008, reported that the spatial sampling biases they identified in Florkart did not affect the leading Isomap components), the temporal non-stationarities in the reference data will need to be further investigated. The finding that Flora Incognita data contains macroecological co-occurrence patterns that match those extracted from Florkart, despite the data limitations in both data sets, indicates that the approach of extracting the leading underlying ecological dimensions minimises the imprint of biases. The known data artefacts and species range shifts have been mostly reported for individual species, i.e. singular dimensions in the high-dimensional binary vector space (Eichenberg et al. 2020). As we have shown here, the few underlying dimensions are related to gradients in climate conditions and soil properties along which species niches are oriented. Because both data sets have been gathered independently, and given that both data sets suffer from their own biases, this emergence of common macroecological patterns is also an opportunity: co-interpreting crowd-sourced data and expert survey unravels robust ecological patterns and benefit from the strengths of both approaches as suggested earlier by Robinson et al. (2020).

Future research into the proposed approach will be oriented along three avenues: 1) novel dimensionality reduction approaches could improve the robustness of the underlying co-occurrence patterns. Although Isomap is much more effective for extracting underlying patterns compared to classical linear alternatives in biogeographical applications (Mahecha et al. 2007, 2009, Van Der Maaten et al. 2012), we expect that also in this area that novel deep-learning algorithms, such as (constrained) variational autoencoders (Kingma et al. 2019) could be used as robust alternatives. 2) An additional improvement for the approach presented here will be exploiting the exact geographical locations of the crowd-sourced data or 3) using the proposed approach for tracking the change in macroecological patterns over time. The latter is possible given the accurate time-information available for each item such that e.g. phenological studies of empirical macroecological patterns are in reach.

## Perspectives for regional to global applications

This study shows that crowd-sourced data from a mobile app do contain substantial macroecological patterns, which in the future could potentially become available at the sub-seasonal time scales and higher spatial resolutions. The next key question that needs to be addressed is whether these advances can be established as a new operational tool for monitoring spatiotemporal macroecological dynamics across continents or even globally. While the technical answer to this question is affirmative, the available apps and data-exploration methods are at a level of maturity where one can even think of assessing the diversity in the hot-spots of diversity, user community engagement will become the critical bottleneck. Undersampling in areas of low population density and the fact that a range of ecologically important species groups tend to be ignored are big challenges. Of course such problems were expected and have been reported for citizen science projects engaged with different species groups (Geldmann et al. 2016, Tiago et al. 2017, Millar et al. 2019, Johnston et al. 2020). As a consequence, crowd-sourced collections as presented here shall not be used for directly estimating levels of species richness and their changes and can only complement expert based biodiversity assessments. For instance, Boersch-Supan et al. (2019) suggested that simple statistical models for estimating population trends from opportunistic lists are robust only for widespread and common species, even in a scheme with many observers and extensive coverage. Data quality limitations of this kind explain the general cautionary reception of citizen science approaches in the scientific literature (Dickinson et al. 2010, Kosmala et al. 2016, Urban et al. 2016, Callaghan et al. 2019b), as well as the ongoing efforts to optimise sampling and evaluation strategies for citizen science data (Specht and Lewandowski 2018, Callaghan et al. 2019a, Kelling et al. 2019).

But even if the current potentials of automated species recognition are not in the direct quantification of species

richness, can such data complement professional inventories or be exploited as inputs to more complex analysis. Under the assumption that apps for automated species identification continue to enjoy broad user uptake and sufficient information comes together (Bonnet et al. 2020), we envisage that analyses like the one presented here could offer novel perspectives for macroecological research. One idea would be to co-interpret the derived with the emerging multivariate satellite derived data cubes that continuously describe the states and processes of land-ecosystems globally (Mahecha et al. 2020). But the most direct way forward is addressing macroecological patterns in time from the seasonal (phenological) time scale to inter-annual dynamics. The latter is needed to quantify the imprint of climate variability on ecosystems as a whole. We hope that automated species identification will become a standard accompanying e.g. coordinated field campaigns.

We expect that the collection of species occurrence data with apps like Flora Incognita or others will soon be used to analyse changes in ecosystems worldwide in near real time. To achieve such a vision, two major challenges need to be addressed: Firstly, data interoperability across different initiatives must be achieved. Secondly, quality assurance and control mechanisms for all crowd-sourced data must be rigorously established. Only if these two challenges are overcome will we be able to tap into the full potential of these new technologies to rapidly assess change in biogeographic dynamics on continental to global scales.

## Conclusions

Almost 1 million georeferenced observations of species occurrences have been collected with a smartphone app during two growing seasons in Germany, which has led to complete coverage of the German traditional inventory grid. Despite of biases in the effective species numbers, this new data collection encodes important macroecological patterns that correspond well to those extracted from the traditional reference database. This finding underscores the potential of smartphone-assisted citizen science and crowd-sourcing for very rapid monitoring of changes in macroecological patterns. Approaches of this kind may complement long-term data collections that explain decadal changes in species composition so far (Blowes et al. 2019). Although this study is regional in scope, it shows that technological advances in the hands of citizen scientists allow monitoring biodiversity transformation in near real time.

*Acknowledgements* – We thank everyone who shared images and geolocations with Flora Incognita, the thousands of volunteers who contributed to the flora of Germany, and the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz) for curating Florkart i.e. Rudolf May. We thank the entire Flora Incognita team for the development of Flora Incognita app, especially David Boho, Hans Christian Wittich and Alice Deggelmann. We thank Fabian Gans, Ingolf Kühn, Bernhard Ahrens, Teja Kattenborn and the subject editor Joaquin Hortal for very helpful discussions and recommendations on the text. Uli Weber helped us in the data preparation phase, and Anke Bebber edited the language. Open access funding enabled and organized by Projekt DEAL.

*Funding* – We thank the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig for supporting this study via a FLEXPOOL project, funded by the German Research Foundation (DFG-FZT 118, 202548816). MR, JW, MS and PM are funded by the German Ministry of Education and Research (BMBF) Grants: 01LC1319A and 01LC1319B; the German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety (BMUB) Grant: 3514685C19, 3519685A08 and 3519685808; Thuringian Ministry for Environment, Energy and Nature Conservation Grant: 68678 and the Stiftung Naturschutz Thüringen (SNT) Grant: SNT-082-248-03/2014.

*Conflict of interests* – Authors declare no competing interests.

### Author contributions

**Miguel D. Mahecha**: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Visualization (equal); Writing – original draft (equal); Writing – review and editing (equal). **Michael Rzanny**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Project administration (equal); Writing – review and editing (equal). **Guido Kraemer**: Formal analysis (equal); Methodology (equal); Writing – review and editing (equal). **Patrick Mäder**: Data curation (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Software (equal); Validation (equal). **Marco Seeland**: Data curation (equal). Jana Wäldchen: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Writing – review and editing (equal).

### Data availability statement

Data and code are available from the Github Digital Repository: <https://github.com/mafla/crowdsourced-ecography-paper> doi: 10.5281/zenodo.4672864 (Mahecha et al. 2021).

## References

Affouard, A. et al. 2017. Pl@ntnet app in the era of deep learning. – In: 5th Int. Conf. on learning representations. ICLR.

Blowes, S. A. et al. 2019. The geography of biodiversity change in marine and terrestrial assemblages. – Science 366: 339–345.

Bodesheim, P. et al. 2018. Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. – Earth Syst. Sci. Data 10: 1327–1365.

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.

Boho, D. et al. 2020. Flora capture: a citizen science application for collecting structured plant observations. – BMC Bioinf. 21: 1–11.

Bonnet, P. et al. 2020. How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. – Ecol. Solut. Evid. 1: e12023.
Bonnet, P. et al. 2018. Plant identification: experts vs. machines in the era of deep learning. – In: Multimedia tools and applications for environmental and biodiversity informatics. Springer, pp. 131–149.
Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.
Brenning, A. 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the r package sperrorest. – In: 2012 IEEE Int. geoscience and remote sensing symposium. IEEE, pp. 5372–5375.
Brondizio, E. et al. 2019. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. – IPBES Secretariat, Bonn.
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.
Dickinson, J. L. et al. 2010. Citizen science as an ecological research tool: challenges and benefits. – Annu. Rev. Ecol. Evol. Syst. 41: 149–172.
Eichenberg, D. et al. 2020. Widespread decline in central european plant diversity across six decades. – Global Change Biol. 27: 1097–1110.
Fick, S. E. and Hijmans, R. J. 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.
Figueroa-Mata, G. and Mata-Montero, E. 2020. Using a convolutional siamese network for image-based plant species identification with small datasets. – Biomimetics 5: 8.
Frey, B. J. and Dueck, D. 2007. Clustering by passing messages between data points. – Science 315: 972–976.
Gaston, K. J. and O'Neill, M. A. 2004. Automated species identification: why not? – Phil. Trans. R. Soc. B 359: 655–667.
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.
Goldsmith, G. R. et al. 2016. Plant-o-matic: a dynamic and mobile guide to all plants of the americas. – Methods Ecol. Evol. 7: 960–965.
Goodfellow, I. et al. 2016. Deep learning. – MIT Press, <www.deeplearningbook.org>.
Hengl, T. et al. 2014. Soilgrids 1 km – global soil information based on automated mapping. – PLoS One 9: e105992.
Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. – Bull. Soc. Vaud. Sci. Nat. 37: 547–579.
Jansen, F. and Dengler, J. 2008. Germansl-a universal taxonomic reference list for phytosociological databases in germany. – Tuexenia 28: 239.
Johnston, A. et al. 2020. Estimating species distributions from spatially biased citizen science data. – Ecol. Model. 422: 108927.
Jones, H. G. 2020. What plant is that? Tests of automated image recognition apps for plant identification on plants from the british flora. – AoB Plants 12: plaa052.
Kelling, S. et al. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. – BioScience 69: 170–179.
Kingma, D. P. et al. 2019. An introduction to variational autoencoders. – Found. Trends Mach. Learn. 12: 307–392.
Kosmala, M. et al. 2016. Assessing data quality in citizen science. – Front. Ecol. Environ. 14: 551–560.
Kraemer, G. et al. 2018. dimred and coranking – unifying dimensionality reduction in r. – R J. 10: 342–358.
Kühn, I. et al. 2004. The flora of german cities is naturally species rich. – Evol. Ecol. Res. 6: 749–764.
Kumar, N. et al. 2012. Leafsnap: a computer vision system for automatic plant species identification. – In: European Conf. on computer vision. Springer, pp. 502–516.
LeCun, Y. et al. 1995. Convolutional networks for images, speech and time series. – In: The handbook of brain theory and neural networks 3361(10). MIT Press.
Legendre, P. and Legendre, L. 2012. Numerical ecology. – Elsevier.
Mäder, P. et al. 2021. The flora incognita app – interactive plant species identification. – Methods Ecol. Evol. doi: 10.1111/2041-210X.13611.
Mahecha, M. D. et al. 2007. Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. – Ecol. Inf. 2: 138–149.
Mahecha, M. D. et al. 2009. Identification of characteristic plant co-occurrences in neotropical secondary montane forests. – J. Plant Ecol. 2: 31–41.
Mahecha, M. D. and Schmidtlein, S. 2008. Revealing biogeographical patterns by nonlinear ordinations and derived anisotropic spatial filters. – Global Ecol. Biogeogr. 17: 284–296.
Mahecha, M. D. et al. 2020. Earth system data cubes unravel global multivariate dynamics. – Earth Syst. Dyn.: 201–234.
Mahecha, M. D. et al. 2021. Data from: Crowd-sourced plant occurrence data provide a reliable description of macroecological gradients. – Github Digital Repository, <https://github.com/mafla/crowdsourced-ecography-paper>, doi: 10.5281/zenodo.4672864.
Meyer, H. et al. 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. – Environ. Model. Softw. 101: 1–9.
Millar, E. E. et al. 2019. The 'cottage effect' in citizen science? Spatial bias in aquatic monitoring programs. – Int. J. Geograph. Inform. Sci. 33: 1612–1632.
Netzwerk Phytodiversität Deutschland and Bundesamt für Naturschutz 2013. Verbreitungsatlas der Farn-und Blütenpflanzen Deutschlands. – Landwirtschaftsverlag.
Nguyen, T. T. N. et al. 2018. Crowdsourcing for botanical data collection towards to automatic plant identification: a review. – Comput. Electron. Agric. 155: 412–425.
Pompe, S. et al. 2008. Climate and land use change impacts on plant distributions in germany. – Biol. Lett. 4: 564–567.
Robinson, O. J. et al. 2020. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. – Divers. Distrib. 26: 976–986.
Rzanny, M. et al. 2019. Flowers, leaves or both? how to obtain suitable images for automated plant identification. – Plant Methods 15: 77.
Specht, H. and Lewandowski, E. 2018. Biased assumptions and oversimplifications in evaluations of citizen science data quality. – Bull. Ecol. Soc. Am. 99: 251–256.
Tenenbaum, J. B. et al. 2000. A global geometric framework for nonlinear dimensionality reduction. – Science 290: 2319–2323.
Terry, J. C. D. et al. 2020. Thinking like a naturalist: enhancing computer vision of citizen science images by harnessing contextual data. – Methods Ecol. Evol. 11: 303–315.

Tiago, P. et al. 2017. Spatial distribution of citizen science casuistic observations for different taxonomic groups. – Sci. Rep. 7: 1–9.
Urban, M. C. et al. 2016. Improving the forecast for biodiversity under climate change. – Science 353(6304): aad8466.
Valavi, R. et al. 2018. blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. – Methods Ecol. Evol. 10: 225–232.
Van Der Maaten, L. et al. 2012. Analyzing floristic inventories with multiple maps. – Ecol. Inf. 9: 1–10.
Van Horn, G. et al. 2018. The inaturalist species classification and detection dataset. – In: Proc. IEEE Conf. on computer vision and pattern recognition, pp. 8769–8778.
Villon, S. et al. 2020. A new method to control error rates in automated species identification with deep learning algorithms. – Sci. Rep. 10: 1–13.
Wäldchen, J. and Mäder, P. 2018a. Machine learning for image based species identification. – Methods Ecol. Evol. 9: 2216–2225.
Wäldchen, J. and Mäder, P. 2018b. Plant species identification using computer vision techniques: a systematic literature review. – Arch. Comput. Methods Eng. 25: 507–543.
Wäldchen, J. et al. 2018. Automated plant species identification – trends and future directions. – PLoS Comput. Biol. 14: e1005993.
Wittich, H. C. et al. 2018. Recommending plant taxa for supporting on-site species identification. – BMC Bioinf. 19: 190.