Research Article

A methodological framework to quantify the spatial quality of biological databases

Jaime R. García Márquez, Carsten F. Dormann, Jan Henning Sommer, Marco Schmidt, Adjima Thiombiano, Sié Sylvestre Da, Cyrille Chatelain, Stefan Dressler & Wilhelm Barthlott

Abstract: The basic information necessary for biogeographical analysis is the geographical location appended to the data contained in biological databases. Reliability of analyses thus crucially depends on the quality of the spatial information available. In the present study we build on a database of vascular plants of West Africa (Ivory Coast, Burkina Faso, Benin), containing 53,205 georeferenced observations distributed over 2,931 collection localities. We propose a methodology to quantify the quality of the database through a series of spatial analyses of spatial configuration of the collection localities, their spatial and environmental bias and inventory completeness. The spatial configuration of the database followed a highly clustered pattern and was strongly biased with respect to the distance to cities, the coast, rivers, roads and protected areas. The same biased pattern was found in relation to several environmental factors. Inventory completeness was calculated by estimating the total number of species based on two non-parametric estimates (first-order Jackknife and Bootstrap) and at different grid cell sizes. At the highest resolution (100 km²) only 5.5% of the cells contained a near-complete (> 80% of Jackknife estimates) species inventory. The percentage of near-complete cells increased as the resolution of analysis decreased. Results of all analyses were integrated into a new index (Gap Selection Index) that serves to guiding future field work campaigns and as cautionary criterion for the uncertainties related to biogeographical application based on the current database.

Keywords: Bootstrap; completeness; environmental bias; Jackknife; multi-scale analysis; point pattern analysis; sampling bias; species richness.

Received: 13 November 2010 – Accepted: 2 November 2011 – Co-ordinating Editor: Florian Jansen.

Introduction

Biogeographical studies aim at understanding how living organisms are spatially distributed, which environmental and biotic parameters influence their distribution and how this pattern changes over time (Brown & Lomolino 1998). The main source of information for such studies is contained in biological databases, specifically lists of species names and their georeferenced locations. Based on this information, spatial biodiversity patterns can be investigated from local to global scales (Brown & Maurer 1989). Therefore, the accuracy of biogeographical analyses heavily depends on the quality of the spatial information recorded in biological databases.

Spatial quality in databases refers specifically to the degree of spatial bias or clustering shown by the location of collection localities (Whittaker et al. 2005), for example towards easily accessible locations (Nelson et al. 1990, Funk & Richardson 2002), conservation areas (Reddy & Dávalos 2003), "diversity hotspots" (Dennis & Thomas 2000) and even the place of residence of biologists (Freitag et al. 1998). This is partly due to the fact that many biological databases are the result of the combination of different heterogeneous data sources (e.g. Küper et al. 2006), each source having its own independent goals and focus areas. Typical information sources are inventories, herbarium and museum collections, atlases and multiple field-based relevés (Zaniewski et al. 2002). One consequence of employing biased data to model the distribution of species or communities might be an erroneous description of real distribution patterns, representing instead the distribution and patterns of sampling effort and/or collection intensity (Williams et al. 2002, Phillips et al. 2009). Likewise, description of species niches could be mis-estimated if the collection records did not sample the whole environmental gradient where a particular species can exist (Raes & ter Steege 2007, Hortal et al. 2008). In addition, biased data can have a negative influence on the performance of predictive modelling techniques (Wolmarans et al. 2010). Given these potential problems, analysis of the amount and nature of geographical bias in a biological database should be an obligatory step when evaluating the quality of biological databases (Romo et al. 2006).

The quality of a biological database can also be evaluated in terms of its floristic completeness. Thereby one can assess how representative the database is in characterizing a specific aspect of biodiversity. Species richness (i.e. the number of species) is the most widely employed index to describe the diversity of an area (Whittaker et al. 2001) and is one of the main criteria to define important areas for conservation (Myers et al. 2000). Hence, decisions for conservation of biodiversity may be inaccurate when based on incomplete information.

In: Dengler, J., Oldeland, J., Jansen, F., Chytrý, M., Ewald, J., Finckh, M., Glöckler, F., Lopez-Gonzalez, G., Peet, R.K., Schaminée, J.H.J. (2012) [Eds.]: Vegetation databases for the 21st century. – Biodiversity & Ecology **4:** 25–39. DOI: 10.7809/b-e.00057.

Table 1: Analysis of species richness and completeness estimates. Richness observed is the total number of species counted in each of the grouping factors. Richness estimates were calculated by using two non-parametric estimation techniques (i.e. First-order Jackknife and Bootstrap). Completeness was calculated by dividing richness observed by richness estimates.

Grouping Factors	Area (km²)	N. Collection localities	Mean Density	Maximum Density	Richness observed	Richness estimates		Completeness	
						Jackknife	Bootstrap	Jackknife	Bootstrap
Study Area	730,600	2,931	0.40	159	4,587	5409	4989	0.85	0.92
Ivory Coast	330,300	876	0.27	7	3,931	4601	4273	0.85	0.92
Burkina Faso	278,800	1,731	0.62	159	1,610	2141	1846	0.75	0.87
Benin	121,500	324	0.27	103	699	854	775	0.82	0.90

It has been shown that the total number of species observed is always less than the true number of species, and hence a negative bias estimator (Walther & Moore 2005: Fig. 2). For example, Palmer (1990) argued that there will always be species present in a sample plot that are not present in the sampled subplots. This may be especially the case for biogeographical studies at regional and even at local scales, where a complete sampling scheme covering the whole study area is impractical (Archaux 2006).

Several different methods exist to estimate the total number of species in a certain area based on a restricted number of samples. Among them, non-parametric techniques (e.g. first- and second-order Jackknife, Bootstrap) have been widely used and have constantly outperformed other techniques, such as species-accumulation curves (e.g. Walther & Martin 2001). By comparing the observed against the estimated number of species, different indices can be calculated to describe the completeness and representativeness of biodiversity information (Soberón et al. 2000; 2007; Soria-Auza & Kessler 2008). One common approach is to stratify the area based on grouping factors and then examine species count completeness in each of them. For example, Parnell et al. (2003) used vegetation classes, forest and non-forest areas, country political divisions and grid cells to identify which areas have received most research effort and therefore possess a more complete biological inventory.

Guidelines for land-use management for plant diversity usually originate from analysis of species distributions and ecosystems health at local scales (Colwell & Coddington 1994). But the scale at which complete information is available generally contrasts with this need. As an example, Soberón et al. (2007), in a study comparing different spatial scales, found the percentage of areas without information to increase with decreasing spatial resolution. Multi-scale analysis may therefore help to identify the scale at which the data is best suited for analysis.

One of the goals of analysing bias, completeness and the effect of spatial scale on biological databases is to answer the questions whether the available information in biological databases is sufficient for the biogeographical research questions at hand, or how much additional effort still needs to be invested, and where.

Over the last nine years, researchers from different institutions and countries have compiled a biological database consisting of georeferenced locations of vascular plants in West Africa. The aim of the present study is to quantify the quality of this biological database in terms of 1) the spatial bias in the distribution of collection localities, 2) the causes or origins of bias in the location of collection localities and 3) the floristic completeness of the database and how it varies at different scales. A final step will be the integration of all these analyses into a Gap Selection Index (GSI) that serves as an identification of areas with missing information and where additional sampling will improve spatial coverage of the database, environmental representativeness and floristic completeness.

Methods

Study area

The study area encompasses 730,600 km² in the countries included as part of the BIOTA project transect in West Africa (i.e. Ivory Coast, Burkina Faso and Benin; Figure 1). The terrain is generally flat with a mean elevation of 277 m a.s.l. However, some mountainous areas in western Ivory Coast reach an altitude of 1,500 m a.s.l.

The study area is characterized by a climatic North-East/South-West gradient. Annual mean temperature ranges from 29.6 °C in the north part of the study area in the Sahelian region to around 18.8 °C in southwestern Ivory Coast. Total annual precipitation shows the opposite gradient: it ranges from 300 mm per year in the North to more than 2600 mm per year in the South-West.

Plant species database

The database used in this study is the result of the compilation of several different heterogeneous sources. Data for Burkina Faso includes vegetation data (Hahn 1996, Kéré 1996, Küppers 1996, Böhm 1998, Denschlag 1998, Ataholo 2001, Krohmer 2004, Schmidt 2006) archived in the West African Vegetation Database (www.westafricanvegetation.org; GIVD-ID AF-00-001; see Janßen et al. 2011, Schmidt et al. 2012) and specimen data from the Herbarium Senckenbergianum (FR) and the Ouagadougou University Herbarium (OUA). This database has been described in detail in Schmidt et al. (2005, 2010a, 2010b). The database covering Ivory Coast is based on herbarium specimens collected since 1900 (described in Aké Assi 2001, 2002) and data collected in the Botanical Garden of Geneva as part of the SIG-Ivoire project (Chatelain et al. 2001). Given the heterogeneity of methodologies and spatial accuracies used to collect the data, the final database was filtered to select the records at a minimal spatial accuracy of 100 km² (10 km × 10 km pixels). At this resolution most of the information from all sources can be utilized for further analysis.

The final database consisted of a total of 53,205 observations distributed over 2,931 collection localities (Fig. 1, Table 1). Collection localities are relevés (field sampling collections) and georeferenced herbarium collections. The data comprise a total of 4,587 plant species belonging to 1,443 genera and 219 families.

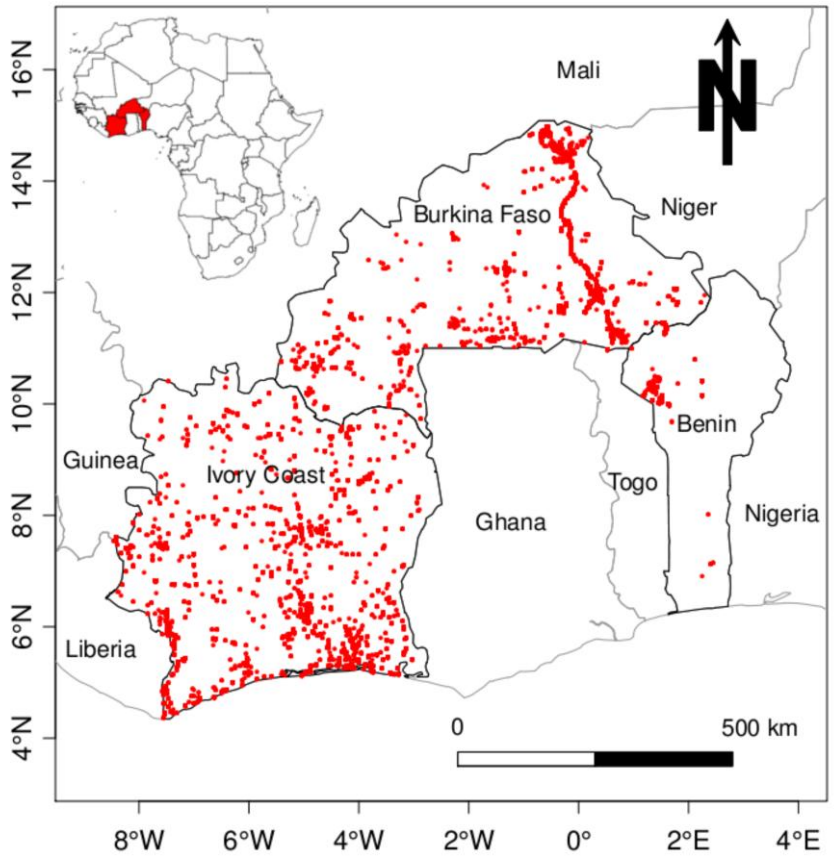

Fig. 1: Detailed map of the study area; red dots represent collection localities.

Environmental data

Table 2 shows the list of environmental data used to evaluate the causes of bias in the data and to check for over- and under-represented environmental conditions. All original layers were prepared and processed using the geographical information system GRASS, Version 6.3 (GRASS Development Team 2008). All layers were transformed to UTM coordinates (zone 30N, datum WGS84), scaled to 100 km² to match the minimal spatial accuracy of the species collections database and clipped to match the study area.

The climatic data were extracted from the WORLDCLIM database (Hijmans et al. 2005). The data were generated through interpolation of average monthly climatic data from weather stations around the world. The elevation layer was also extracted from the WORLDCLIM database and was included into the geographical information system SAGA (SAGA Development Team 2008) to derive the wetness index variable. Elevation variance was computed by calculating the variance of the elevation values using a 9 × 9 cell moving window.

Statistical analysis

All analysis were carried out using the statistical software R (R Development Core Team 2009), with package vegan (Oksanen et al. 2009) and spatstat (Baddeley & Turner 2005).

Density estimates and departure from complete spatial randomness (CSR)

Density estimates and departure from randomness of collection localities was investigated from the theoretical backgrounds of point pattern analysis (see Chapter 8 in Cressie 1993). Collection localities were considered as the "points" used in point pattern analysis. The first step was to calculate the density as the number of collection localities per 100 km².

To visualise density patterns, a density map of the study area was created using an isotropic Gaussian kernel (Diggle 2003, Baddeley & Turner 2005). The bandwidth of the Gaussian kernel was estimated using the method of Berman & Diggle (1989) which minimises the Mean Square Error (MSE) of the kernel estimator (see Appendix 1). 30 km was chosen as the final bandwidth, although other values seem plausible given the flatness of the curve. Locality density is one of the inputs for the Gap Selection Index (GSI) (see below).

To quantitatively test whether the distribution pattern of the collection localities departed from a complete spatial random distribution (CSR, henceforth), Ripley's *K*-function was used (Schabenberger & Gotway 2005: pp. 99–103), following the procedures implemented in Baddeley et al. (2000). Point-wise envelops under CSR were computed based on 100 simulations of random distributed points over the study area. Then, it was checked whether the observed pattern (i.e. the one defined by the collection localities) lay inside this envelope.

Bias analysis

The purpose of the bias analysis was three-fold: (1) to understand which factors cause spatial bias in the distribution of collection localities; (2) to check whether spatial bias of collection localities represents environmental bias as well; and (3) to generate a layer representing environmental bias in the study area. The detail procedures carried out for each of the above points are described below.

Procedure 1: To measure the magnitude of bias in collection localities, each of the bias factors (see Table 3) was split into four intervals based on the range of measured distances. Thus, interval 1 represented the area where distances to each bias factor were smallest, while in interval 4 distances were highest. To calculate the size of each interval the Fisher algorithm was used (Fisher 1958). This method selects class breaks to group similar values and at the same time maximizes the difference between classes (Slocum et al. 2005).

Next, bias was quantified for each interval following the index of Kadmon et al. (2004):

$$Bias_d = \frac{n_d - p_d N}{\sqrt{p_d (1 - p_d) N}} \quad (1.1)$$

where n_d is the number of collection localities within a specified interval (d), N is the total number of collection localities in the database and p_d is the probability for a given collection locality to be within a interval (d). Since the above equation is derived from the normal approximation to the binomial distribution, values become statistically significant when they are greater or less than 1.64 and –1.64 respectively (at $\alpha = 0.05$). Bias values greater

than 1.64 represent over-sampled areas, that is areas with more collection localities than expected from a random sampling design. In contrast, bias values less than –1.64 depicted under-sampled areas.To estimate *p* for each interval, the same amount of points as collection localities was generated based on a random sampling design with replacement. The fraction of random points within each interval was taken to be *p*. The definition of random points and the estimation of the bias index was repeated 100 times. Basic statistics and confidence intervals were calculated.

Table 2: List of bias factors and environmental data used to evaluate the sources of spatial bias and the environmental representativeness in the distribution of collection localities. Distance to the coast has a different meaning for Burkina Faso: it represents possible bias in a north-south gradient within the country.

Layer name	Derived layer name	Abbreviation	Source
Bias Factors			
Main cities	Distance to cities		DMA (1992)
Countries of the world	Distance to the coast		DMA (1992)
Rivers	Distance to rivers		DMA 1992)
Roads	Distance to roads		DMA 1992)
World database on protected areas	Protected areas		World Conservation Union & UNEP-World Conservation Monitoring Centre (2007)
Environmental layers			
Annual mean temperature	Annual mean temperature	amte	Hijmans et al. (2005)
Annual precipitation	Annual precipitation	apre	Hijmans et al. (2005)
Temperature annual range	Temperature annual range	tara	Hijmans et al. (2005)
Elevation	Elevation	elev	Hijmans et al. (2005)
Elevation	Elevation variance of elevation	srtm	Hijmans et al. (2005)
Elevation	Wetness index	weti	Hijmans et al. (2005)

Table 3: Differences between the number and percentage of grid cells containing information at different spatial resolutions.

Resolution (km^2)		Total No. of cells	Cells with some information	%
100	(10 km x 10 km)	7306	1011	13.8
900	(30 km x 30 km)	884	440	48.2
3,660	(60 km x 60 km)	247	182	73.7
14,400	(120 km x 120 km)	74	63	85.1

Procedure 2: Even if collection localities are biased towards some of the bias factors considered here, applying predictive modelling may still be valid as long as the geographical arrangement of those bias factors properly represent the environmental variability of the study area. To assess whether localities covered environmental conditions randomly, several steps were carried out. First, the bias factors that showed over-representation of collection localities in any of the four intervals were selected. Second, the number of collection localities present in the selected bias factors in the specified interval was counted and the same number of points was created randomly throughout the study area. Third, both sets of points were overlaid with the environmental layers described above in order to obtain the values of the environmental variables for each point. Fourth, the frequency distribution of those values was compared using the Kolmogorov-Smirnov test (KS). The KS tests the null hypothesis that the frequency distribution of two samples were drawn from the same continuous distribution (Marsaglia et al. 2003).

Procedure 3: A new layer representing the environmental bias in the study area was created following the same steps as in procedure 1 but using the environmental layers instead of the bias factors. Once the bias index was calculated for each environmental layer and for each interval, all layers were summed up to derive the environmental bias index map. This layer was used as input for the Gap Selection Index.

Database completeness

To analyse the floristic completeness of the database used in this study, the completeness index proposed by Soberón et al. (2000) was used. This index is based on the comparison of the total (i.e. estimated) number of species present in a certain geographical area (*S**) with the number of species observed (S_{obs}) in the same area: $C = S_{obs}/S*$ where *C* is the completeness index. The calculation of the *C*-index has to be constrained to a certain geographical area or subdivisions of it, called grouping factor herein. In this study the *C*-index was calculated for the whole study area, for each country and for grid cells of different size to identify how the completeness of the database varies with scale.

The observed number of species (S_{obs}) in each grouping factor was the number of species counted. To estimate the total number of species (*S**) two non-parametric techniques were implemented:

1. First-order Jackknife as a bias reduction method:

$$S* = S_{obs} + L\frac{n-1}{n}$$

where n is the number of samples and L the number of species that occur in only one sample (Burnham & Overton 1979, Heltshe & Forrester 1983).

2. Bootstrap:

$$S* = S_{obs} + \sum (1 - p_i)^N$$

where p_i is the frequency of species i and N is the total number of collections in the grouping factor (Smith & Belle 1984).

Database quality evaluation

We developed the Gap Selection Index as a measure of database quality. For that we considered three factors: the density of collection localities as calculated using the Gaussian smooth kernel (d), the values representing the environmental bias in each country (b) and the database completeness (C). All factors were converted to values between 0 and 1 following the equation of Legendre and Legendre (1998):

$$y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad (1.2)$$

Then, all factors were subtracted from 1 to ensure that values close to 1 represent deficiencies in data quality. The gap selection index was thus calculated as:

$$GSI = \frac{3 - d' - b' - C'}{F} \quad (1.3)$$

where F represent the number of factors included in the index. Results of the index are values between 0 and 1, where values close to zero represent areas that have been properly represented while values close to one represent areas where the density of collection is very low or zero, the information is incomplete and where the environmental conditions are not well represented in the distribution of collection localities.

Results

Density and complete spatial randomness

The mean density of collection localities in the study area is very low, with less than 1 collection locality per 100 km² (Table 1). Collection localities are unevenly distributed, with certain patches of high densities, especially in Burkina Faso and Benin (Fig. 2a). This clustered pattern was quantitatively estimated based on the analysis of the inhomogeneous K-function (Fig. 2b).

Bias analysis

In general, all bias factors had a strong influence on the spatial distribution of collection localities in the study area (Fig. 3). For Ivory Coast, there was a clear over-representation of collection localities in areas close to each of the bias factors (i.e. interval 1), but most importantly to cities, to the coast and to roads. In contrast, in distant areas the trend was towards an under-representation of collection localities (Fig. 3). Closeness to roads and specifically to protected areas were the factors explaining the over-representation of collection localities in Burkina Faso. Also in Burkina Faso there seemed to be a preference to collect far away from the main cities, the coast (which in this case represented north-south gradient) and roads (Fig. 3). As for Benin, in places situated close to rivers and roads, an over-representation of collection localities was found while at long distances a negative bias existed (Fig. 3).

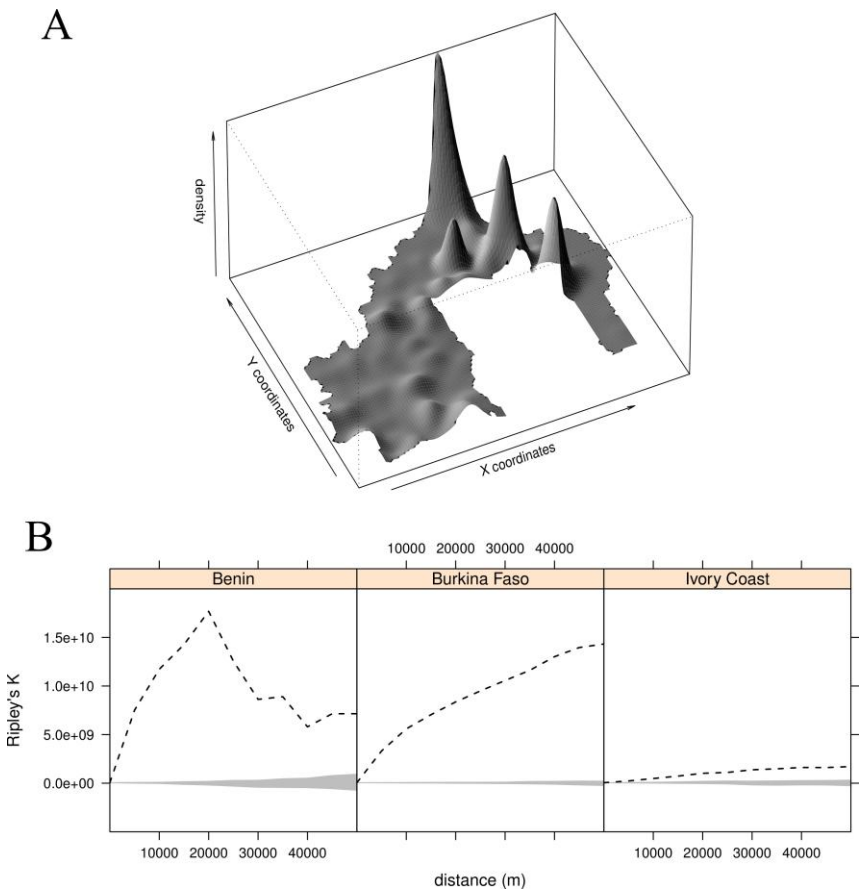

Fig. 2: (A) Three dimensional view of collection localities density patterns estimated based on a smoothing Gaussian kernel; (B) Point pattern estimates of the collection localities based on the inhomogeneous Ripley's *K*-function. Displayed are envelops (gray) representing the area ocuppied by realizations of 100 simulated random patterns. Black dash lines are the estimated *K* values of the collection localities for different distances. The line is expected to be inside the envelop if the pattern of collection localities is random. Lines above the envelope indicate a clustered pattern.

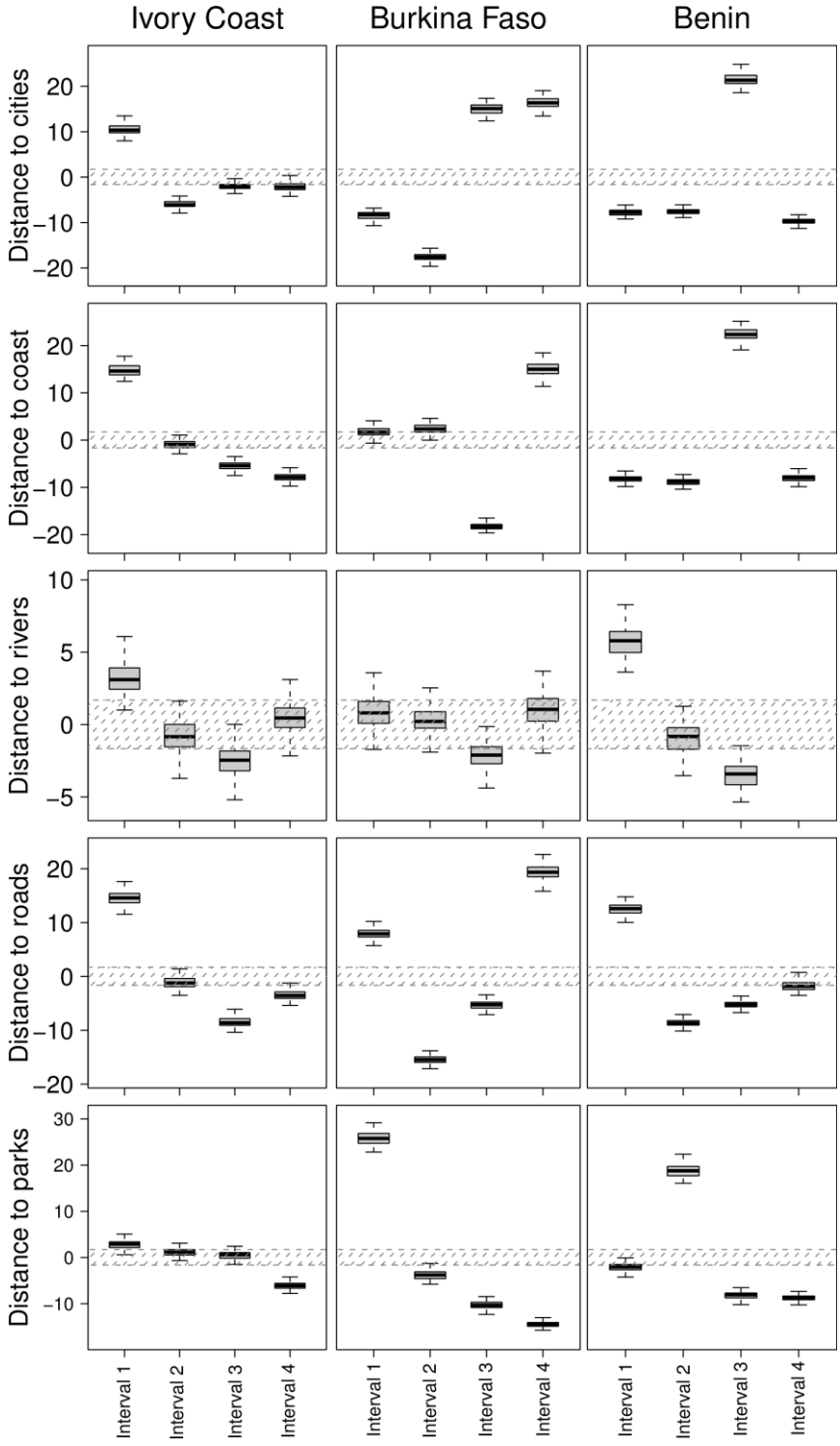

Fig. 3: Bias estimates (as calculated from equation 1.1) for each of the bias factors (rows) in each country (columns) and for each distance interval (1 to 4) considered in this study. Interval 1 represents short distance and interval 4 largest distance values. Shadow polygons represent the range of values where no bias is expected. If boxplots are within this area than the number of collection localities are as expected from a random sampling scheme (i.e. no bias). Boxplots above and below this area represent over- or under-sampling, respectively.

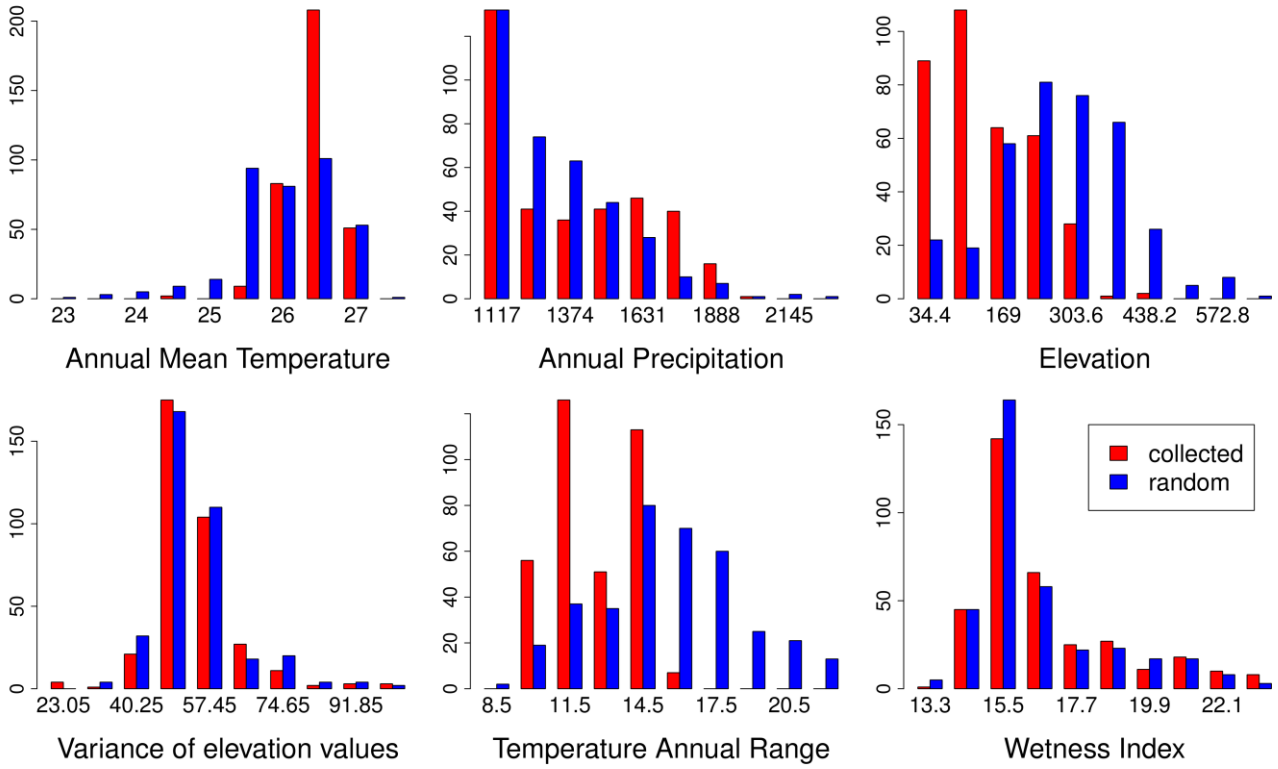

Fig. 4: Example of the difference between the frequency distribution of environmental values found for all collections located in areas close to cities (i.e. interval 1) in Ivory Coast and for the locations of randomly distributed points in the study area. No significant differences exists for the variance of elevation (srtm) and wetness index (weti). On the contrary, for all other environmental variables the differences are significant (see also Appendix 2). There is an over-representation of low elevation areas while areas of high altitude have been under-represented. The same case applies for temperature annual range (tara) and the opposite for annual mean temperature (amte) and annual precipitation (apre).

In general, an over-representation of collection localities in some regions of the study area also correlated with an environmental bias. That means, some environmental condition were over-represented (e.g. those found near to roads), while others are under-represented (those found far away from parks) in the distribution of collection localities. Few of these biases were consistent, however. For example, in Ivory Coast, environmental conditions far away from cities were representatively sampled, while in Burkina Faso they were over-, and in Benin under-represented (Fig. 3, first row, interval 4). Significant differences between sampled versus randomised locations were found for the frequencies of the values of all environmental parameters (see Fig. 4 for an example from interval 1; see also Appendix 2). If the distribution of collection localities was not environmentally biased, one would have expected to find no differences between these frequencies. An exception was Ivory Coast for elevational variance (srtm) and wetness index (weti), where no bias was present despite the fact that the majority of collection localities were near cities and rivers.

The map in Figure 5 depicts the sum of the bias estimates for each of the environmental variables used in this study. Clearly, environmental conditions in coastal Ivory Coast, in and around the eastern Guinean forest in Benin and the Sahelian zone in Burkina Faso were over-represented. In contrast, wide expanses of savannas and forest-savanna mosaic in all three countries were under-represented.

Completeness analysis

A general comparison between the two non-parametric techniques employed indicated that results of the first-order Jackknife estimator were in general higher than results of the Bootstrap estimator and therefore completeness values were always higher when calculated based on the Bootstrap technique.

Estimates of species richness and completeness were calculated for different grouping factors (Table 2). In general, the floristic knowledge of the study area was good, as shown by the high values of the completeness index. Comparing the three countries independently, Burkina Faso is the least studied country since it has the lowest completeness value. From 1,610 plant species observed at the time, there will be at least several hundred species still not described in the database.

Completeness analysis was also applied on a grid cell basis. Different cell sizes (i.e. resolutions) were used (i.e. 100 km², 900 km², 3,600 km², 14,400 km²). Correlations between number of species observed and estimated were very high at all resolutions (in all cases a correlation coefficient of 0.99). In contrast, correlations between estimated species richness and completeness values were low (Fig. 6). In general, grid cells with the highest number of species were not necessarily complete. Complete cells occurred in Benin and Burkina Faso, although the two countries were less studied than Ivory Coast (Table 1). Note that virtually all of Benin had a completeness index close or equal zero.

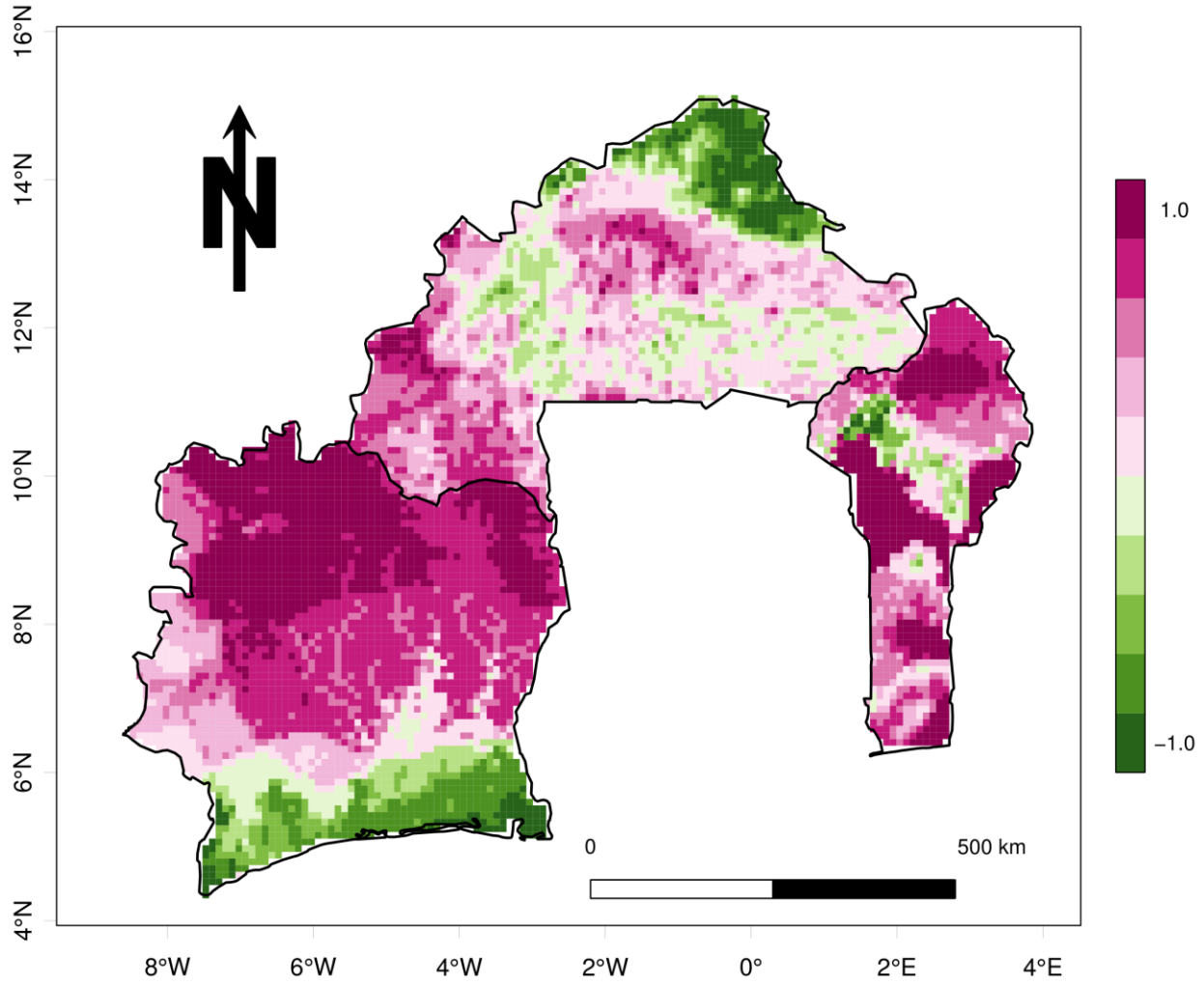

Fig. 5: Environmental bias in the study area. Values close to one (red) represent areas where environmental conditions are under-represented. Areas assigned values close to zero (white) have been visited as expected by applying a random sampling scheme. Environmentally over-represented areas in the distribution of collections localities are those with values close to –1 (green).

The percentage of grid cells containing information increased with an increase in cell size (Table 3). As a result of the clustered distribution pattern of collection localities, there were few areas with high density and most of the remnant area had either no or a very small density of collection localities. Consequently there are either areas with high completeness index values and areas with very low completeness values. However, the percentage of grid cells with a completeness value equal or higher than 0.6 increased up to a grid cell size of 3,600 km². There is also a constant increase of grid cells with completeness values higher than 0.8 as the resolution increases (Fig. 7).

Gap selection index

Based on the density distribution of collection localities (Fig. 2b), the degree of environmental bias (Fig. 5) and the floristic completeness of the database (Fig. 6), the gap selection index was calculated (Fig. 8). In this index, values close to zero represent areas that have been well studied, where the density of collection localities is high, where the environmental conditions have been properly represented **and** where the floristic information is complete.

From a pessimistic point of view, problematic intervals can be considered as those having values greater than 0.8. 71.1% of the total area were within this interval. 70.9%, 64.9% and 86.2% of the area in Ivory Coast, Burkina Faso and Benin, respectively, had values greater or equal 0.8.

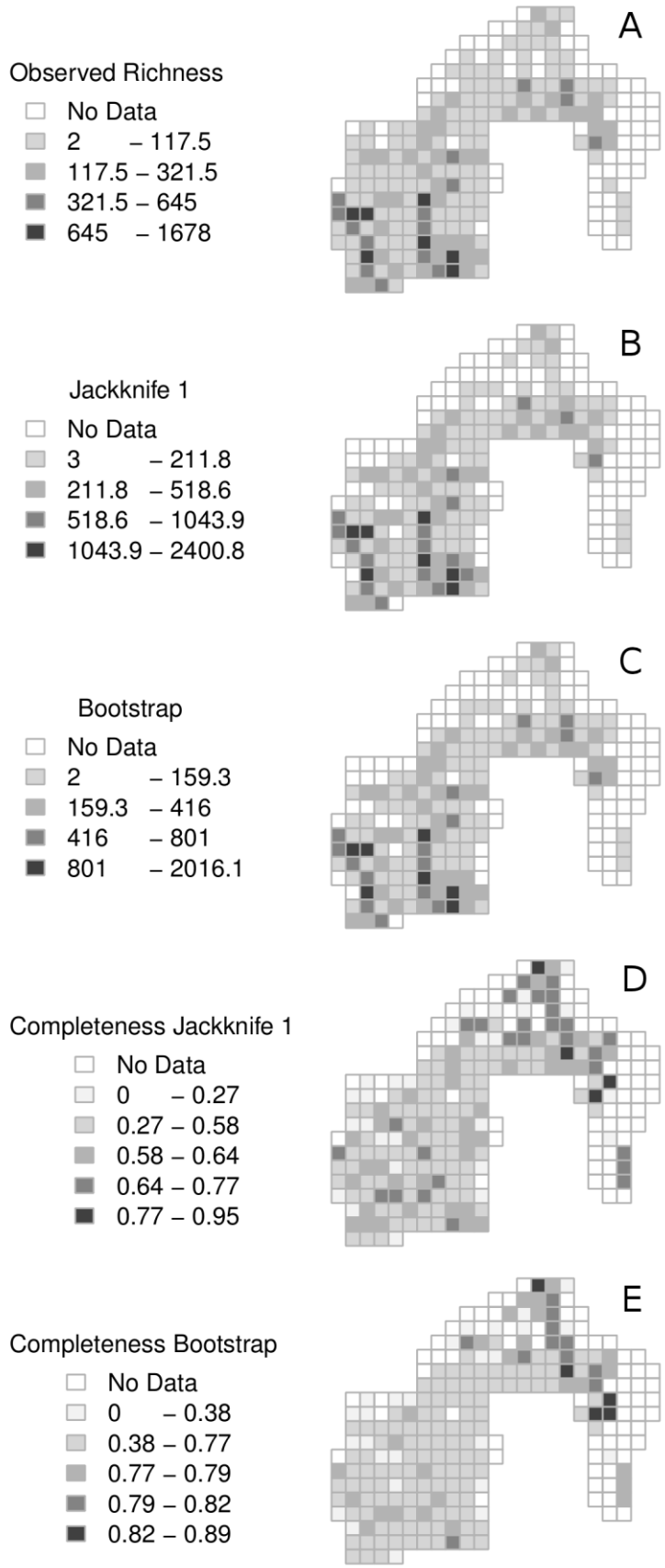

Fig. 6: Maps of observed species richness (A) and estimated species richness as calculated using two non-parametric estimation techniques (i.e. First-order Jackknife (B) and Bootstrap (C)). In the fourth and fifth rows are the maps of the Completeness Index (D and E) (i.e. richness observed divided by richness estimates). All illustrations are based on the analysis done at a 3,600-km² resolution. The first-order Jackknife estimator produced in all cases higher species richness estimates while the Bootstrap produced more conservative numbers.

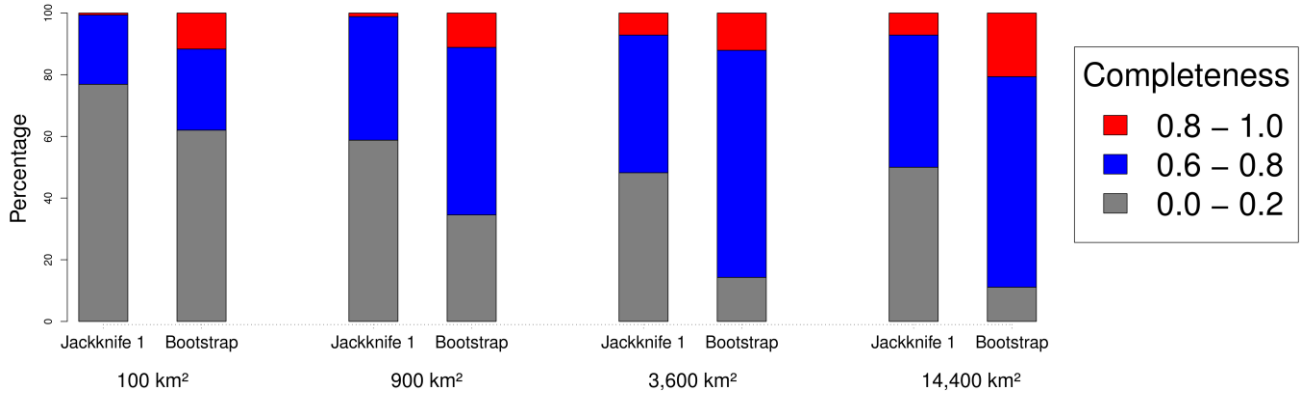

Fig. 7: Barplots showing the percentage of grid cells in three completeness classes (see legend). Calculations were done for four different spatial resolutions and considering the two non-parametric techniques used for species richness estimations (i.e. First-order Jackknife and Bootstrap) at each resolution.

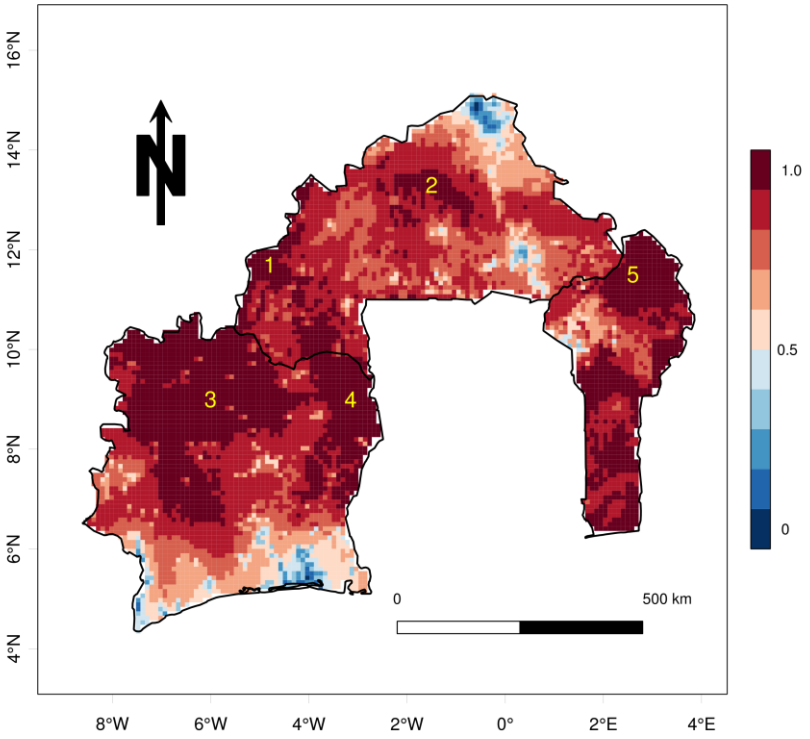

Fig. 8: Map of the Gap Selection Index (GSI). The main goal of the index is to emphasize those areas that have been poorly visited and contained environmental information not well represented by the distribution of collection localities in the database. Therefore, values close to one represent under-represented areas while places with values close to zero have received enough attention and have been well studied. The index has been calculated by integrating information on collection densities, environmental representativeness and floristic completeness on a pixel based approach (i.e. 100 km²). Regions with highest GSI values are marked in yellow numbers: (1) Triangle Fô/Bobo-Dioulasso/Samorogouan, close to border with Mali: area of rice fields, traditional Vitellaria parklands, some forested hills in its northern part (often holy groves, thus protected); (2) hill chain of the Lake Bam/Lake Dem/Kaya area with a low cover of woody plants, but including some rare species, e.g. *Boswellia dalzielii;* (3) area between Mt. Sanghe NP, Ferkessédougou and the Mali border; (4) area around the eastern border of Comoé-NP; (5) northern-most part of Benin including W National Park with highest values in the Atakora hunting zone close to the border with Burkina Faso: Sudanian savannas and parklands.

Discussion

Uneven efforts of plant collection in West Africa

The database used in this study is the result of more than nine years of compilation efforts. Still, the average density of collection localities is very small (Table 1) and spatially strongly clustered (Fig. 2). Although the database has already been used for estimating patterns of plant diversity in the region (Schmidt et al. 2005, Thiombiano et al. 2006), this was not the main goal motivating the construction of the database. Instead, many of the data have been generated from specific projects with their focus on specific areas and research questions. In Burkina Faso, for example, a special focus has been given to the Sahelian acacia savanna ecoregion, where, in consequence, collection localities are found at high densities. A specific macroecological analysis of this particular region was presented by Schmidt et al. (2008). Other "hot-spots" of plant collections in Burkina Faso are situated in small areas that have been the focus of investigation by students and researches in the region (Müller 2003, Krohmer 2004). The same situation occurs in Benin, where only a special area has been the focus of research, namely the Atacora mountains around Natitingou (Sieglstetter 2002, Krohmer 2004), where the maximal density found is 103 collections per 100 km² (Fig. 2a).

Not surprisingly, the south of Ivory Coast has been studied better, since in this area the Guinean Forest diversity hotspot is located (Myers et al. 2000), one of the known areas with high species richness and endemism. Several studies have been carried out to investigate different aspects of the composition, structure and dynamics of the forest ecosystems in this area (Chatelain et al. 2004, Nussbaumer et al. 2005).

Bias: a recurrent issue

Spatial bias in biological databases is one of the most repeatedly mentioned issues in biogeographical research. It is presumed to be one of the factors potentially distorting the results of biogeographical analysis (Funk & Richardson 2002, Loiselle et al. 2008, Wolmarans et al. 2010), but its influence on model output is rarely explicitly made. However, it has been demonstrated that spatial bias can have a substantial influence on model outcome and performance as well as in the establishment of the effect of environmental variables on the defined niche of a species (Graham et al. 2004, Feeley & Silman 2010). For that reason, we concur with other authors that before using predictive modelling techniques, it is necessary to explicitly evaluate the database in terms of spatial bias and to understand the possible causes that led to that bias.

We used a similar approach to identify factors influencing spatial bias in the database and estimating environmental bias as implemented earlier by Kadmon et al. (2004) and Loiselle et al. (2008). Although Kadmon et al. (2004) found significant differences between the distribution of collection localities and that of the rainfall conditions based on a random selection of localities in the study area, they demonstrated that predictions of habitat suitability were not biased, since the statistical difference was weak (although significant). In this study, strong statistical differences were found and therefore we conclude that model predictions based on the current database are likely to produce biased and misleading estimates of species range predictions.

What is the appropriate scale of analysis?

Finding the appropriate scale of analysis is one of the most controversial and studied issues in ecology (Hurlbert & Jetz 2007). The resolution of analysis should comply with the inherent properties of any given dataset (Hengl 2006). At the same time, it should be adequate to solve the ecological questions of concern. One of the future applications of the database presented in this study is to generate species distribution patterns taking advantage of the high spatial accuracy of the data (i.e. 100 km²) and the accessibility of environmental information available at this resolution. Is then 100-km² resolution the proper scale for such studies?

Several of the analysis carried out in this study indicate otherwise. Hengl (2006) recommended the inspection of the density of a point pattern as one of the criteria to define the right pixel size. Results show that the mean square error of the bandwidth calculated to estimate the density (Appendix 1) is higher at 10 km and diminishes at longer distances with small differences above 30 km. That means that if we calculate the density patterns with a bandwidth of 10 km, the final estimates will have a bigger error than at longer distances.

Secondly, it is clear from Table 3 and Figure 7 that the percentage of grid cells containing information increases as the resolution increases and that the amount of grid cells with complete information also increases as the cell size increases. In conclusion, more accurate models of species distribution patterns can be obtained if a cell size bigger than 100 km² is used. We suggest a cell size of 3,600 km² (i.e. 60 km × 60 km) as a minimum acceptable scale of analysis.

Disentangling the gap selection index

It is not the first time that a methodology was developed to identify areas where information is missing (see Küper et al. 2006 for an example). Funk et al. (2005) developed a method which they called survey-gap analysis to identify the location of future collection activities. For that purpose, they used a set of environmental variables to derive an environmental diversity (ED) measure (see Faith & Walker 1996) and a set of collected sites, which they integrated into a complementary analysis to select sites that would contribute new taxa. The novelty of the Gap Selection Index concept developed in this study is the integration of different independent criteria, which makes the selection of target sites objective and efficient. Relying on only one criterion makes the identification of target sites impractical. For example, if some particular places are to be visited based on density estimates then areas with low density of collection localities will be chosen. Obviously, information is still missing in those areas. But if the environmental conditions characterizing those areas are very similar to areas where collection density is high, then similar vegetation structure and floristic composition can be expected and no novel data will be added to the database. A more efficient use of the resources at hand will be to go to sites that combine low collection densities and underrepresented environmental conditions.

Another important combination of criteria for site selection uses collection densities and database completeness, compared on a grid cell basis (Soberón et al. 2007). The expected behaviour of the relationship between these two criteria is an increase in completeness with an increase in collection density. In the study, this relationship is weak. In general, the majority of grid cells has low density values and is yet complete. In conclusion, a good estimation of the floristic composition of

the study area requires few collection localities, but properly distributed.

One of the main purposes of the Gap Selection Index is to support the prioritisation of future sampling efforts to complement current inventory data of the floristic composition of the region. Large areas are unsampled still and visiting all of them is unrealistic. The use of additional tools that help in the effective selection of target areas is needed. One of these tools is GoogleEarth, where the Gap Selection Index can be displayed and compared with high resolution images. Visualisation of this type could help both understanding the potentially underlying drivers of diversity and identifying areas that deserve more investment regarding fieldwork and resources. There are already efforts in making use of the integration of GIS analysis and visualisation on GoogleEarth for scientific and communication purposes (e.g. Conroy et al. 2008).

In this study we decided to use non-parametric techniques (i.e. first order Jackknife and Bootstrap) to estimate species richness based on the species observed in the different collection localities. The first-order Jackknife one has been consistently ranked among the most precise techniques and the second one is considered as a technique that generally underestimates the real value (Beck & Schwanghart 2010). The two techniques were selected so that a range of possible values could be given and compared. However, there is a range of species richness estimator techniques that have been used for the same purpose, all likely to underestimate species richness except in near-complete inventories (Coddington et al. 2009, Unterseher et al. 2011). For example, Baselga & Novoa (2006) and Jiménez et al. (2009) used rarefaction curves to estimate species richness values and compared them to the observed species richness to evaluate the completeness of their databases.

Modelling plant diversity in West Africa: where from now?

Is it possible to use the information in the database to model plant diversity patterns in West Africa? From the map of the Gap Selection Index (Fig. 8) it is clear that a considerable part of the study area is missing information and is not well represented in the database. Although several techniques can potentially be applied to model diversity patterns, it is important to recognize that there will be a considerable amount of uncertainty present in predictions, especially in those areas missing information. It is recommended that any efforts to estimate and display plant diversity patterns should be accompanied by the Gap Selection Index map as a representation of the uncertainties of the outcomes.

Many techniques can be used to predict and model plant diversity patterns based on our database. These techniques can be grouped into two main modelling approaches. In the first group are all techniques that directly relate species diversity (e.g. species richness) and environmental variables. From the relationships found (i.e. variable coefficients), models are able to predict species diversity to the total extent of the region of interest.

To employ this approach for data with limited completeness one would select only those grid cells that are above a selected completeness threshold for modelling (e.g. Romo et al. 2006). Although information might thus be lost, the reliability of predictions increases since the modelling itself is based on more accurate data. Another possibility is to use all cells with data, so no information is lost, and weight those cells with the completeness index calculated for each of them.

The second group of modelling approaches is the species niche modelling (see Elith et al. 2006, for a review and performance comparison of different methods). The principle of this approach is to model each species individually and create species range map for each of them. Afterward, all maps are stacked up to create a final map of species richness. A major constraint of applying this approach to the database is that most species have been collected only very few times. If the recommendation of using species with 10 or more sampled occurrences given by Hernández et al. (2006) is followed, than only 1,423 species will be considered for analysis, that is, only 31% of all species. However, even as few as three occurrence localities have claimed to be useful to model species ranges (Pearson et al. 2007).

Dealing with spatial bias (i.e. spatial autocorrelation) in the distribution of the collection localities becomes an issue when using any of these approaches (Dormann 2007). Some research has been done to deal with this issue (e.g. Kadmon et al. 2004, Allouche et al. 2008, De Marco et al. 2008, Phillips et al. 2009) and the methodologies recommended can potentially be also applied for the database (see Dormann et al. 2007 for a review of methods to deal with spatial autocorrelation). Algar et al. (2009) made a comparison of the two main approaches described above to estimate species richness patterns and predictions to the future and found that after dealing with the spatial autocorrelation issue the first approach (i. e. empirical diversity theory approaches) fared significantly better.

Conclusions

A lot of work, money and effort has been invested in the creation of the databases forming the basis for this study. Scientific investigations have already set a good example for utilizing these data for research and conservation applications (Schmidt et al. 2005, 2008, Thiombiano et al. 2006). However, the use of the database for macroecological studies at regional scales might be limited by a series of factors. The distribution of collection localities has not been done in a manner expected in statistical techniques. Particularly, collections do not follow a random distribution in geographic as well as in environmental space but rather a very clustered pattern. There are few areas and environments that have been well investigated but information is still missing for most of the extent of the study area and its habitats.

The correlation between several bias factors and the distribution of collection localities is very high. Unfortunately, this collection distribution bias represents environmental bias as well. Many areas with specific environmental conditions have not been visited yet, and their inclusion into models that seek to predict species distribution ranges to those areas will result in misleading estimates.

If biogeographical applications based on the database are expected in the short term, it is strongly recommended to find the proper modelling techniques that are robust to spatial bias in the distribution of collection localities. Several approaches and modelling techniques have been tested to deal with this issue (see, e.g., Kadmon et al. 2004, Phillips et al. 2009) with positive results. Nonetheless, if there are funds and resources to organize field campaigns, targeting the areas identified by the Gap Selection Index will fill up data gaps and will decrease the amount of bias currently present in the database.

Acknowledgements

We thank the German Ministry of Education and Research for financing the BIOTA project (funding code: 01LC0617D1), all collectors for the herbaria OUA and FR, as well as data contributors to the vegetation database. C.F.D. was funded by the Helmholtz Association (VH-NG 247). The authors also acknowledge funding from the Hessian Initiative for the Development of Scientific and Economic Excellence (LOEWE) through the Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main.

References

Aké Assi, L. (2001): Flore de la Côte d'Ivoire: catalogue systématique, biogéographique et écologie. I – Boissiera **57**: 1–396.

Aké Assi, L. (2002): Flore de la Côte d'Ivoire: catalogue systématique, biogéographique et écologie. II – Boissiera **58**: 1–401.

Algar, A.C., Kharouba, H.M., Young, E.R., Kerr, J.T. (2009): Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. – Ecography **32**: 22–33. CrossRef

Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A., Kadmon, R. (2008): Incorporating distance constraints into species distribution models. – Journal of Applied Ecology **45**: 599–609. CrossRef

Archaux, F., Gosselin, F., Bergès, L., Chevalier, R. (2006): Effects of sampling time, species richness and observer on the exhaustiveness of plant censuses. – Journal of Vegetation Science **17**: 299-306. CrossRef

Ataholo, M. (2001): Pflanzensoziologische Untersuchungen der Segetalarten in der Sudanzone Westafrikas. – PhD thesis, J.W. Goethe-Universität, Frankfurt am Main [Deposited at the University Library of Frankfurt].

Baddeley, A., Möller, J., Waagepetersen, R. (2000): Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. – Statistica Neerlandica **54**: 329–350. CrossRef

Baddeley, A., Turner, R. (2005): Spatstat: an R package for analyzing spatial point patterns. – Journal of Statistical Software **12(6)**: 1–42.

Baselga, A., Novoa, F. (2006): Diversity of *Chrysomelidae* (*Coleoptera*) in Galicia, northwest Spain: Estimating the completeness of the regional inventory. – Biodiversity and Conservation **15**: 205–230. CrossRef

Beck, J., Schwanghart, W. (2010): Comparing measures of species diversity from incomplete inventories: an update. – Methods in Ecology and Evolution **1**: 38–44. CrossRef

Berman, M., Diggle, J. (1989): Estimating weighted integrals of the second-order intensity of a spatial point process. – Journal of the Royal Statistical Society **51**: 81–92.

Böhm, M. (1998): Dorfvegetation in Burkina Faso. – Diploma thesis, J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].

Brown, J.H., Lomolino, M.V. (1998): Biogeography. 2nd ed. – Sunderland: Sinauer.

Brown, J.H., Maurer, B.A. (1989): Macroecology: The division of food and space among species on continents. – Science **243**: 1145–1150. CrossRef

Burnham, K.P., Overton, W.S. (1979): Robust estimation of population size when capture probabilities vary among animals. – Ecology **60**: 927–936. CrossRef

Chatelain, C., Dao, H., Gautier, L., Spichiger, R. (2004): Forest cover changes in Côte d'Ivoire and Upper Guinea. – In: Poorter, L., Bongers, F., Kouamé, F.N., Hawthorne, W.D. [Eds.]: Biodiversity of West African forests. An ecological atlas of woody Plant species: 15–32. Wallinford: CABI Publisher.

Chatelain, C., Gautier, L., Spichiger, R., (2001): Application du SIG Ivoire à la distribution potentielle des espèces en fonction des facteurs écologiques. – Systematics and Geography of Plants **71**: 313–326. CrossRef

Coddington, J.A, Agnarsson, I., Miller, J.A., Kuntner, M., Hormiga, G. (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. – Journal of Animal Ecology **78**: 573–584. CrossRef

Colwell, R.K., Coddington, J.A. (1994): Estimating terrestrial biodiversity through extrapolation. – Philosophical Transactions of the Royal Society of London, B **345**: 101–118. CrossRef

Conroy, G.C., Anemone, R.L., Regenmorter, J.V., Addison, A. (2008): Google Earth, GIS, and the great divide: A new and simple method for sharing paleontological data. – Journal of Human Evolution **55**: 751–755. CrossRef

Cressie, N.A. (1993): Statistics for spatial data. – New York. J. Wiley.

De Marco, P. jr., Felizola Diniz-Filho, J.A., Bini, L.M. (2008): Spatial analysis improves species distribution modelling during range expansion. – Biology Letters 4(5):577-80 CrossRef

Dennis, R.L.H. & Thomas, C.D. (2000): Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. – Journal of Insect Conservation **4**: 73–77 CrossRef

Denschlag, J. (1998): Ethnobotanische und pflanzensoziologische Untersuchungen der Gehölzvegetation bei den FulBe im Südosten von Burkina Faso (Westafrika). – Diploma thesis, J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].

Diggle, P. (2003): Statistical analysis of spatial point patterns. 2nd ed. – London: Arnold Publishers.

DMA [Defense Mapping Agency] (1992): Digital chart of the world. – Fairfax, Virginia: Defense Mapping Agency.

Dormann, C.F. (2007): Effects of incorporating spatial autocorrelation into the analysis of species distribution data. – Global Ecology and Biogeography **16**: 129–138. CrossRef

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

Elith, J., Graham, C.H., Anderson, R.P., Dudìk, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E. (2006): Novel methods improve prediction of species' distributions from occurrence data. – Ecography **29**: 129–151. CrossRef

Faith, D., Walker, P., 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. – Biodiversity and Conservation **5**: 399–415. CrossRef

Fisher, W.D., 1958. On grouping for maximum homogeneity. – Journal of the American Statistical Association **53**: 789–798. CrossRef

Freitag, S., Hobson, C., Biggs, H.C., Jaarsveld, A.S.V., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. – Animal Conservation **1**: 119–127.

Funk, V., Richardson, K., 2002. Systematic data in biodiversity studies: use it or lose it. – Systematic Biology **51**: 303–16. CrossRef

Funk, V., Richardson, K.S., Ferrier, S. (2005): Survey-gap analysis in expeditionary research: where do we go from here. – Biological Journal of the Linnean Society **85**: 549–567. CrossRef

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T. (2004): New developments in museum-based informatics and applications in biodiversity

analysis. – Trends in Ecology and Evolution. **19**: 497–503. CrossRef
GRASS Development Team (2008): Geographic Resources Analysis Support System (GRASS GIS) Software. – Open Source Geospatial Foundation. URL: http://grass.osgeo.org.
Hahn, K. (1996): Die Pflanzengesellschaften der Savannen im Südosten Burkina Fasos (Westafrika). Ihre Beeinflussung durch den Menschen und die naturräumlichen Gegebenheiten. – PhD thesis,J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Heltshe, J., Forrester, N. (1983): Estimating species richness using the jackknife procedure. – Biometrics **39**: 1–11. CrossRef
Hengl, T. (2006): Finding the right pixel size. – Computers & Geosciences **32**: 1283–1298. CrossRef
Hernández, P.A., Graham, C.H., Master, L.L., Albert, D.L. (2006): The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography **29**: 773–785.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. (2005): Very high resolution interpolated climate surfaces for global land areas. – International Journal of Climatology **25**: 1965–1978. CrossRef
Hortal, J., Jimenez-Valverde, A., Gomez, J.F., Lobo, J.M., Baselga, A. (2008): Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos **117**: 847–858. CrossRef
Hurlbert, A.H., Jetz, W. (2007): Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. – Proceedings of the National Academy of Sciences of the USA **104**: 13384–13389. CrossRef
Janßen, T., Schmidt, M., Dressler, S., Hahn-Hadjali, K., Hien, M., Konaté, S., Lykke, A.M., Mahamane, A., Sambou, B., Sinsin, B., Thiombiano, A., Wittig, R., Zizka, G. (2011) Addressing data property rights concerns and providing incentives for collaborative data pooling: the West African Vegetation Database approach. – Journal of Vegetation Science **22**: 614–620.
Jiménez, I., Distler, T., Jørgensen, P. M., (2009): Estimated plant richness pattern across northwest South America provides similar support for the species-energy and spatial heterogeneity hypotheses. – Ecography **32**: 433–448.
Kadmon, R., Farben, F., Danin, A. (2004): Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecological Applications **14**: 401–413. CrossRef
Kéré, U. (1996): Die Dorf- und Savannenvegetation in der Region Tenkodogo (Burkina Faso). – PhD thesis, J.W. Goethe-Universität, Frankfurt am Main. [Deposited at the University Library of Frankfurt].
Kreft, H., Jetz, W. (2007): Global patterns and determinants of vascular plant diversity. – Proceedings of the National Academy of Sciences of the USA **104**: 5925–5930. CrossRef
Krohmer, J. (2004): Umweltwahrnehmung und -klassifikation bei Fulbegruppen in verschiedenen Naturräumen Burkina Fasos und Benins. – PhD thesis,J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Küper, W., Sommer, J.H., Lovett, J., Barthlott, W. (2006): Deficiency in African plant distribution data-missing pieces of the puzzle. – Botanical Journal of the Linnean Society **150**: 355–368.
Küppers, K. (1996): Die Vegetation der Châine de Gobnangou. – PhD thesis, J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Legendre, P., Legendre, L. (1998): Numerical Ecology. 2nd ed. – Amsterdam: Elsevier Science.
Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G., Montiel, O.M. (2008): Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – Journal of Biogeography **35**: 105–116.
Marsaglia, G., Tsang, W.W., Wang, J. (2003): Evaluating Kolmogorov's distribution. – Journal of Statistical Software **8(18)**: 1–4.
Müller, J.V. (2003): Zur Vegetationsökologie der Savannenlandschaften im Sahel Burkina Fasos. – PhD thesis,J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J. (2000): Biodiversity hotspots for conservation priorities. – Nature **403**: 853–858. CrossRef
Nelson, B.W., Ferreira, C.A.C., da Silva, M.F., Kawasaki, M.L. (1990): Endemism centers, refugia and botanical collection density in Brazilian Amazonia. – Nature **345**: 714–716. CrossRef
Nussbaumer, L., Gautier, L., Chatelain, C., Spichiger, R. (2005): Structure et composition floristique de la forêt classée du Scio, Côte d'Ivoire. Etude descriptive et comparative. – Candollea **60**: 393–502.
Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G., Solymos, P., Stevens M.H.H., Wagner, H. (2010). vegan: community ecology package. R package version 1.17-4.
Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., D'Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnut, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., (2001). Terrestrial ecoregions of the world: A new map of life on earth. – BioScience **51**: 933 – 938. CrossRef
Palmer, M.W. (1990): The estimation of species richness by extrapolation. – Ecology **71**: 1195–1198. CrossRef
Parnell, J.A.N., Simpson, D.A., Moat, J., Kirkup, D.W., Chantaranothai, P., Boyce, P.C., Bygrave, P., Dransield, S., Jebb, M.H. P., Macklin, J., Meade, C., Middleton, D.J., Musaya, A.M., Prajaksood, A., Pendry, C.A., Pooma, R., Suddee, S., Wilkin, P. (2003): Plant collecting spread and densities: their potential impact on biogeographical studies in thailand. – Journal of Biogeography **30**: 193–209. CrossRef
Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend Peterson, A. (2007): Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – Journal of Biogeography **34**: 102–117. CrossRef
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., (2009): Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecological Applications **19**: 181–197. CrossRef
R Development Core Team (2009): R: A Language and Environment for Statistical Computing. – Vienna: R Foundation for Statistical Computing. URL: http://www.R-project.org.
Raes, N., ter Steege, H. (2007): A null-model for significance testing of presence-only species distribution models. – Ecography **30**: 727–736. CrossRef
Reddy, S., Dávalos, L.M. (2003): Geographical sampling bias and its implication for conservation priorities in Africa. Journal of Biogeography **30**: 1719–1727. CrossRef
Romo, H., García-Barros, E., Lobo, J.M. (2006): Identifying recorder-induced geographic bias in an Iberian butterfly database. – Ecography **29**: 873–885. CrossRef
SAGA Development Team (2008): System for Automated Geoscientific Analyses (SAGA GIS). – URL: http://www.saga-gis.org/.
Schabenberger, O., Gotway, C. (2005): Statistical methods for spatial data analysis. – London: Chapman & Hall.
Schmidt, M. (2006): Pflanzenvielfalt in Burkina Faso — Analyse, Modellierung und Dokumentation. – PhD thesis,J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Schmidt, M., König, K., Müller, J. (2008): Modelling species richness and life form composition in sahelian Burkina Faso with remote sensing data. – Journal of

Arid Environments **72**: 1506–1517. CrossRef
Schmidt, M., Kreft, H., Thiombiano, A., Zizka, G. (2005): Herbarium collections and field data-based plant diversity maps for Burkina Faso. – Diversity and Distributions **11**: 509–516. CrossRef
Schmidt, M., Thiombiano, A., Ouédraogo, A., Hahn-Hadjali, K., Dressler, S., Zizka, G. (2010a): Assessment of the flora of Burkina Faso. – In: van der Burgt, X., van der Maesen, J., Onana, J.-M. [Eds.]: Systematics and conservation of African plants: 571–576. Kew: Royal Botanic Gardens.
Schmidt, M., Thiombiano, A., Ouédraogo, A., Hahn-Hadjali, K., Dressler, S., Zizka, G. (2010b): Phytodiversity data – strengths and weaknesses, a comparison of collection and relevé data from Burkina Faso. – In: van der Burgt, X., van der Maesen, J., Onana, J.-M. [Eds.]: Systematics and conservation of African plants: 571–576. Kew: Royal Botanic Gardens.
Schmidt, M., Janßen, T., Dressler, S., Hahn, K., Hien, M., Konaté, S., Lykke, A.M., Mahamane, A., Sambou, B., Sinsin, B., Thiombiano, A., Wittig, R., Zizka, G. (2012): The West African Vegetation Database. – In: Dengler, J., Oldeland, J., Jansen, F., Chytrý, M., Ewald, J., Finckh, M., Glöckler, F., Lopez-Gonzalez, G., Peet, R.K., Schaminée, J.H.J. [Eds.]: Vegetation databases for the 21st century. – Biodiversity & Ecology **4**: 105–110. Hamburg: Biocentre Klein Flottbek and Botanical Garden. CrossRef
Sieglstetter, R. (2002): Wie die Haare der Erde — Vegetationsökologische und soziokulturelle Untersuchungen zur Savannenvegetation der Südsudanzone Westafrikas und ihrer Nutzung und Wahrnehmung durch die ländliche Bevölkerung am Beispiel der Region Atakora im Nordwesten Benins. – PhD thesis,J.W. Goethe University Frankfurt, Frankfurt am Main [Deposited at the University Library of Frankfurt].
Slocum, T., McMaster, R., Kessler, F., Howard, H. (2005): Thematic cartography and geographic visualization. – Upper Saddle River, NJ: Prentice Hall.
Smith, E.P., von Belle, G. (1984): Nonparametric estimation of species richness. – Biometrics **40**: 119–129.
Soberón, J., Jiménez, R., Golubov, J., Koleff, P. (2007): Assessing completeness of biodiversity databases at different spatial scales. – Ecography **30**: 152–160.
Soberón, J.M., Llorente, J.B., Oñate, L. (2000): The use of specimen-label databases for conservation purposes: an example using Mexican papilionid and pierid butterflies. – Biodiversity and Conservation **9**: 1441–1466. CrossRef
Soria-Auza, R., Kessler, M. (2008): The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. – Diversity and Distributions **14**: 123–130. CrossRef
Stockwell, D.R.B., Peterson, A.T. (2002): Effects of sample size on accuracy of species distribution models. – Ecological Modelling **148**: 1–13. CrossRef
Thiombiano, A., Schmidt, M., Kreft, H., Guinko, S. (2006): Influence du gradient climatique sur la distribution des espèces de *Combretaceae* au Burkina Faso (Afrique de l'Ouest). – Candollea **61**: 189–213.
Unterseher, M., Jumpponen, A., Öpik, M., Tedersoo, L., Moora, M., Dormann, C.F., Schnittler, M. (2011) Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. – Molecular Ecology **20**: 275–285. CrossRef
Walther, B.A., Martin, J.L. (2001): Species richness estimation of bird communities: how to control for sampling effort? – Ibis **143**: 413–419. CrossRef
Walther, B.A., Moore, J.L. (2005): The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. – Ecography **28**: 815–829. CrossRef
Whittaker, R.J., Araujo, M.B., Paul, J., Ladle, R.J., Watson, J.E.M., Willis, K.J. (2005): Conservation biogeography: assessment and prospect. – Diversity and Distributions **11**: 3–23. CrossRef
Whittaker, R.J., Willis, K.J., Field, R. (2001): Scale and species richness: towards a general, hierarchical theory of species diversity. – Journal of Biogeography **28**: 453–470. CrossRef
Williams, P.H., Margules, C.R., Hilbert, D.W. (2002): Data requirements and data sources for biodiversity priority area selection. – Journal of Bioscience **27**: 327–338. CrossRef
Wolmarans, R., Robertson, M.P., van Rensburg, B.J. (2010): Predictive invasive alien plants distributions: how geographical bias in occurrence records influences model performance. – Journal of Biogeography **37**: 1797–1810. CrossRef
World Conservation Union, UNEP-World Conservation Monitoring Centre (2007): World database on protected areas. – Cambridge, UK: WCMC.
Zaniewski, A.E., Lehmann, A., Overton, J. (2002): Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. – Ecological Modelling **157**: 261–280.

Jaime R. García Márquez*
(jrgarcia.marquez@gmail.com)
Conservation International Colombia
Carrera 13 N. 71–41
Bogotá, COLOMBIA

Carsten F. Dormann
(carsten.dormann@biom.uni-freiburg.de)
Biometry & Environmental System Analysis, Faculty of Forest and Environmental Science, University of Freiburg
Tennenbacher Str. 4
79106 Freiburg, GERMANY

Jaime R. García Márquez, Jan Henning Sommer (hsommer@uni-bonn.de), Sié Sylvestre Da (sylda@uni-bonn.de) & Wilhelm Barthlott (unb12b@uni-bonn.de)
Nees Institute for Biodiversity of Plants, University of Bonn
Meckenheimer Allee 170
53115 Bonn, GERMANY

Jaime R. García Márquez & Marco Schmidt
(marco.schmidt@senckenberg.de)
Biodiversity and Climate Research Center (BiK-F)
Senckenberganlage 25
60325 Frankfurt am Main, GERMANY

Adjima Thiombiano
(adjima_thiombiano@univ-ouaga.bf)
University of Ouagadougou
Post box : 09 BP
848 Ouagadougou 09, UFR/SVT,
BURKINA FASO

Cyrille Chatelain (cyrille.chatelain@ville-ge.ch)
Conservatoire et Jardin botaniques
Ch. de l'Impératrice 1 CP 60
1292 Chambésy, Genève,
SWITZERLAND

Marco Schmidt & Stefan Dressler
(stefan.dressler@senckenberg.de)
Dept. of Botany and Molecular Evolution, Senckenberg Research Institute
Senckenberganlage 25
60325 Frankfurt, GERMANY

Jan Henning Sommer
Department of Ecology and Resource Management, ZEF Center for Development Research, University of Bonn
Walter-Flex-Str. 3
53113 Bonn, GERMANY

Marco Schmidt & Stefan Dressler
J.W. Goethe University Frankfurt,
Institute for Ecology, Evolution and Diversity
Max-von-Laue-Str. 9
60438 Frankfurt, GERMANY

*Corresponding author