Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records

Article *in* Ecography · June 2018

DOI: 10.1111/ecog.03944

3 authors, including:

Sophie Monsarrat
Aarhus University
25 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Graham I. H. Kerley
Nelson Mandela University
297 PUBLICATIONS **7,038** CITATIONS
SEE PROFILE

ECOGRAPHY

Research

Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records

Sophie Monsarrat, Andre F. Boshoff† and Graham I. H. Kerley

S. Monsarrat (https://orcid.org/0000-0002-6220-5306) (sophiemonsarrat@gmail.com), A. F. Boshoff and G. I. H. Kerley, Centre for African Conservation Ecology, Nelson Mandela Univ., Port Elizabeth, South Africa.

Ecography
41: 1–12, 2018
doi: 10.1111/ecog.03944

Subject Editor: Joaquin Hortal
Editor-in-Chief: Miguel Araújo
Accepted 23 June 2018

Historical biodiversity occurrence records are often discarded in spatial modeling analyses because of a lack of a method to quantify their sampling bias. Here we propose a new approach for predicting sampling bias in historical written records of occurrence, using a South African example as proof of concept. We modelled and mapped accessibility of the study area as the mean of proximity to freshwater and European settlements. We tested the model's ability to predict the location of historical biodiversity records from a dataset of 2612 large mammal occurrence records collected from historical written sources in South Africa in the period 1497–1920. We investigated temporal, spatial and environmental biases in these historical records and examined if the model prediction and occurrence dataset share similar environmental bias. We find a good agreement between the accessibility map and the distribution of sampling effort in the early historical period in South Africa. Environmental biases in the empirical data are identified, showing a preference for lower maximum temperature of the warmest month, higher mean monthly precipitation, higher net primary productivity and less arid biomes than expected by a uniform use of the study area. We find that the model prediction shares similar environmental bias as the empirical data. Accessibility maps, built with very simple statistical rules and in the absence of empirical data, can thus predict the spatial and environmental biases observed in historical biodiversity occurrence records. We recommend that this approach be used as a tool to estimate sampling bias in small datasets of occurrence and to improve the use of these data in spatial analyses in ecological and conservation studies.

Keywords: citizen-science, environmental bias, historical ecology, mammals, occurrence records, South Africa

Introduction

With a growing interest in historical ecology research (Szabó 2015) and the recognition that historical biodiversity data are key to understanding the long-term impact of human activities (Clavero and Revilla 2014, Mihoub et al. 2017), more and more datasets of centuries-old biodiversity records are being assembled (Rookmaaker 2007, Matthews and Heath 2008, Clavero and Delibes 2013, Butynski et al. 2015,

†Deceased

NORDIC SOCIETY OIKOS
www.ecography.org

Boshoff et al. 2016, Turvey et al. 2015, 2017). However, an obstacle to integrating historical occurrences in ecological analyses remains the substantial levels of spatial and environmental bias these data contain due to opportunistic and unstandardized sampling and reporting. Existing methods to address sampling bias in spatial modeling analyses are not appropriate for small single-taxa datasets of occurrence. The lack of appropriate methodological approaches to quantify sampling bias in these data often leads to the discarding of historical occurrence records altogether, along with the valuable information they contain (Szabó and Hédl 2011). It is thus critical to develop methods to quantify sampling effort of historical written records and explicitly incorporate sampling biases in spatial analyses in order to allow their consideration in ecology and conservation research.

Historical sources such as governmental archives, travel accounts, gazetteers, diaries and correspondence, contain a wealth of information on the distribution and abundance of plants and animals, long before modern ecological data started to be collected. By extending the timeline considered, written historical records of species occurrence represent an opportunity to contribute to a better understanding of biodiversity trends over long periods of times (Shaffer et al. 1998, Tingley and Beissinger 2009, Turvey et al. 2017). They are also important in a conservation context, as they can provide unique new insights into extinction dynamics and changing species status through time, help to establish meaningful temporal baselines for biodiversity, and assist in determining desired future conditions, all of which are key to setting conservation priorities and informing management decisions (Willis et al. 2007, Rick and Lockwood 2013, Turvey et al. 2015, Mihoub et al. 2017).

However, historical data are often perceived as untrustworthy and assumed to be inadequate for most statistical modeling methods. Notably, the lack of information on non-detection and the absence of protocols for data collection make their utilization in spatial modeling techniques difficult (Syfert et al. 2013). Ideally, sampling effort should be perfectly uniform or random so that observed distribution patterns are real and not simply a reflection of the intensity of sampling. However, processes behind the collection of historical records are often spatially biased towards regions more frequented by observers (Reddy and Davalos 2003, Newbold 2010). If this spatial bias results in an environmental bias, this may induce a bias towards environments that have received more sampling and substantially impact quantitative analyses based on these data. Spatial modeling techniques where an empirical model relates such occurrence records to environmental variables will then likely provide inaccurate outputs, reflecting sampling effort rather than the true distribution of the species (Phillips et al. 2009). Ultimately, the inability to measure sampling bias may lead to rejecting these data, as they may not achieve minimum requirement to be used in conservation analyses (Williams et al. 2002). Tools to explicitly report the spatial distribution of the bias and lack of sampling effort across a study region include maps of ignorance that provide information on sampling coverage and reliability (Rocchini et al. 2011, Ruete 2015), maps of collecting effort (Schulman et al. 2007) or spatial modeling of the distribution of effort based on occurrence data from biodiversity databases and environmental variables that influence where observers are likely to search for particular species (Stolar and Nielsen 2015). Solutions for explicitly incorporating sampling bias in species distribution modeling include 1) the spatial filtering of occurrence data (Boria et al. 2014, Fourcade et al. 2014, Varela et al. 2014), 2) weighting sample points according to the distribution of sampling effort (Stolar and Nielsen 2015), 3) the manipulation of background data (also referred to as pseudo-absences) using all occurrences within a target group as absence data with the hypothesis that these share the same geographical bias as the presence dataset (Phillips et al. 2009, Hertzog et al. 2014, Ranc et al. 2016) or 4) the incorporation of presence-only and presence–absence data for multiple species in a joint probabilistic model to estimate and adjust for the bias (Fithian et al. 2015). These methods have been shown to improve the performance of SDMs built with spatially biased data. Fourcade et al (2014) demonstrated that systematic sampling, or the spatial filtering of occurrence data, consistently ranked among the best performing method to correct for sampling bias in the widely used species distribution modelling tool MAXENT. However, because they require large datasets of a species' occurrence or information on the occurrence of other species collected with the same protocol, these methods are inappropriate for application to occurrence record datasets that have a small sample size or focus on one taxa, which is often the case of datasets extracted from historical written sources (Hoving et al. 2003, Matthews and Heath 2008, Kittinger et al. 2013). Providing an appropriate method to predict sampling bias in historical occurrence records would represent an important step towards integrating these data in spatial analyses of biodiversity patterns.

A South African example

Over-hunting and loss of habitat largely altered the composition of the large mammal fauna in southern Africa, especially since the start of the colonial period (Boshoff and Kerley 2015, Boshoff et al. 2016). However, because most of this impact occurred in the past 250–300 years, these trends cannot be captured by research studies based on recent ecological data alone. Since the first written account of South Africa's fauna by Vasco de Gama in 1497, literate explorers, settlers, hunters, missionaries and naturalists have written accounts describing their environment and the animals they encountered in their travels. Occurrence records extracted from these written sources provide invaluable insights into the historical distribution, abundance and composition of the mammal fauna in South Africa (Rookmaaker 1989, Skead et al. 2007, 2011, Scholte 2012, Boshoff and Kerley 2013). Previous studies have discussed the reliability of these records (Boshoff and Kerley 2010) and their implication for the historical

distribution of large mammals in South Africa (Boshoff et al. 2002, Boshoff and Kerley 2015, Boshoff et al. 2016), but the sampling biases have not been explicitly quantified. Evaluating the sampling bias in these records will allow their inclusion in more advanced quantitative analyses, with the potential to reconstruct the historical distribution, abundance and extinction patterns of large mammals in Southern Africa.

Here, we propose an innovative approach, requiring no empirical data, to map modelled geographical accessibility, based on our knowledge of the behavior and environmental constraints faced by observers. We test the relevance of these accessibility maps to predict sampling bias in empirical data, using a comprehensive dataset of written records of species' occurrence collected in the early historical period in South Africa that we assume to be representative of the sampling effort at this period. We investigate temporal, spatial and environmental biases in these historical records to inform further use of these data for ecology and conservation research. Finally, we examine if the model predictions and the occurrence dataset share similar environmental bias, compared to what would be expected given a uniform use of the study area.

Material and methods

Study area

The boundaries of the study area follow those described in Boshoff et al. (2016). It incorporates the present-day political territories of the Western Cape, Eastern Cape, Northern Cape and Free State provinces, and the far western part of the North West Province, of the Republic of South Africa, and all of the Kingdom of Lesotho (Fig. 1). The area constitutes some 70% (881 377 km^2) of the total area of 'South Africa', i.e. South Africa and the countries of Lesotho and Swaziland.

Historical occurrence records

We used a data set of 3512 historical occurrence records of medium- to large-sized mammals, including 51 species from 11 families (comprising 16 carnivores and 39 herbivores). Each occurrence record corresponds to sightings, vocalisations or signs (e.g. tracks/spoor) of one or more individuals. This data set was extracted from written historical sources including letters, diaries or books written by various missionaries, explorers, travellers, naturalists, military personnel, big game hunters and pastoralists who visited or settled in the study area, starting in 1497, this being the start of the written historical period in that region (Skead 1980, 1987, Boshoff and Kerley 2013) (Supplementary material Appendix 1). Because these historical occurrence records concern a large number of mammal species, over a long period of time, we consider these data to be representative of the true distribution of observers in the early historical period, defined here as ending in 1920, after which period the study of the distribution of mammals in South Africa became more formalized and started to be captured in the scientific literature (Boshoff and Kerley, 2013). We deleted duplicates, i.e. records reporting occurrences of different mammal species but collected at the same place at the same date by the same observer. After this step, 2612 unique localities were left for the analyses. We tested if the occurrence records are spatially biased using simulation envelopes, a well-established technique to test

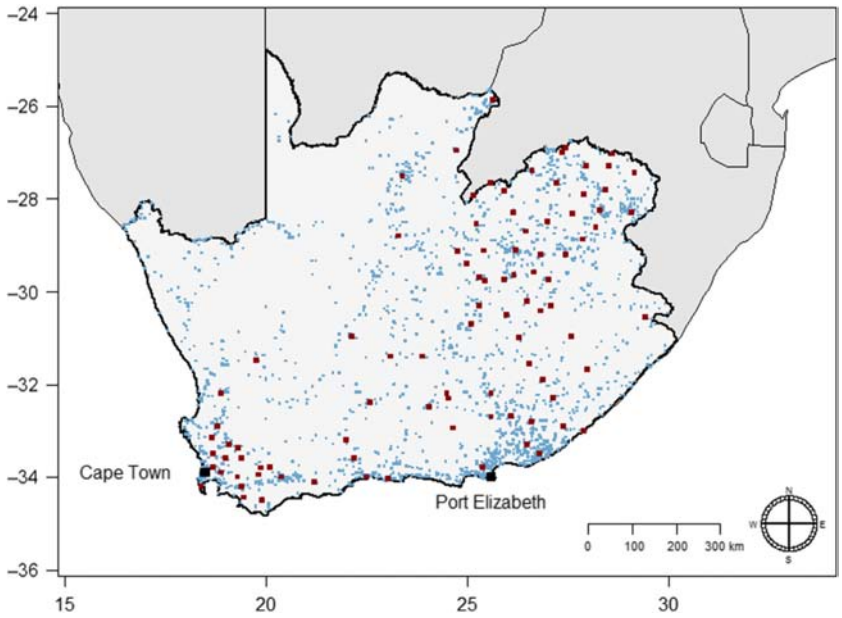

Figure 1. Map of South Africa showing the study area (in light grey) and locality of historical written records (blue dots). Locations of historical European settlements (established pre-1900) in the study area are indicated with red squares and the location of Cape Town and Port Elizabeth are indicated with black squares, for geographical reference purpose.

complete spatial randomness (CSR) (Baddeley et al. 2014). This method is based on computing a summary function of the point pattern, such as Ripley's K function (Dixon 2002), and comparing it with the envelope of the same functions obtained from several simulations of the null model, in that case a homogeneous Poisson process (Supplementary material Appendix 2). We also mapped the distribution of historical records through time to identify temporal biases in the data, for six time periods: pre-1720 and five 40-year periods between 1720 and 1920 (Supplementary material Appendix 3).

Environmental biases in the historical occurrence records

We tested environmental biases in the historical written records by comparing the frequency distribution in the historical written occurrence dataset (OBS) and the background dataset (BACKGROUND, defined using the coordinates of all cells in the study area, n = 8387) for three environmental features: maximum temperature of the warmest month (Tmax), mean monthly precipitation (PREC) and mean annual net primary productivity (NPP). We selected these variables because they have been identified as correlates of large mammals' distribution and species richness in southern Africa (Coe et al. 1976, Andrews and O'Brien 2000) and would potentially be useful predictors of species distribution in habitat modelling approaches based on this occurrence dataset. We applied a Mann-Whitney U test to determine whether the distributions of the two datasets along these three gradients are identical. We calculated the difference between the OBS and BACKGROUND frequencies, as an index of environmental 'preference' of the observers (black line on Fig. 2). An index above (below) 0 indicates a preference (avoidance) of the environmental condition considered. Finally, we calculated the environmental completeness of the OBS dataset, i.e. the degree to which the climatic ranges are covered by the observations, following the methodology in Kadmon et al. (2003).

Maximum temperature of the warmest month Tmax and average monthly precipitations were downloaded from WorldClim 1.4 at a 30 s (~1 km²) resolution (Hijmans et al. 2005). PREC was calculated as the mean of the average precipitation for the 12 months of the year. NPP was obtained from MODIS MOD17 gross/net primary production project of the numerical terradynamic simulation group, at 30 s resolution (Zhao et al. 2005). We reduced the impact of inter-annual climatic variability by using long-term climatology obtained by averaging Tmax and PREC over the period 1960–1990 and NPP, which relies on more recent satellite data, over the period 2000–2015. By measuring environmental biases in 15th–20th century data with environmental data from the 20th century, we assumed that environmental conditions have remained stable during that time period. Environmental data were aggregated on a 0.1° × 0.1° grid using the bilinear interpolation resampling method, with the raster package (Hijmans 2014) in R 3.4.3 (R Core Team).

Accessibility model

We expected that early observers would be biased in their movement by the proximity of freshwater and existing settlements (Supplementary material Appendix 1). Thus the accessibility model was built based on two spatial components: the proximity of European settlements and freshwater. We decided not to include terrain or barriers (e.g. cliffs, rivers) in the accessibility model, owing to the peculiar behavior of early travelers who demonstrated a strong ability to overcome obstacles in the landscape (see Supplementary material Appendix 4 for more details on the observer's habits and the effect of adding terrain in the analysis). Important European settlements of the 17th–19th century in the study area were identified based on Floyd's chronological order of town establishment in South Africa (Floyd 1960). We retrieved information on surface freshwater from the 1:500 000 Resource Quality Information Services river coverage dataset provided by the Department of Water and Sanitation of the Republic of South Africa (Weepener et al. 2012). These data

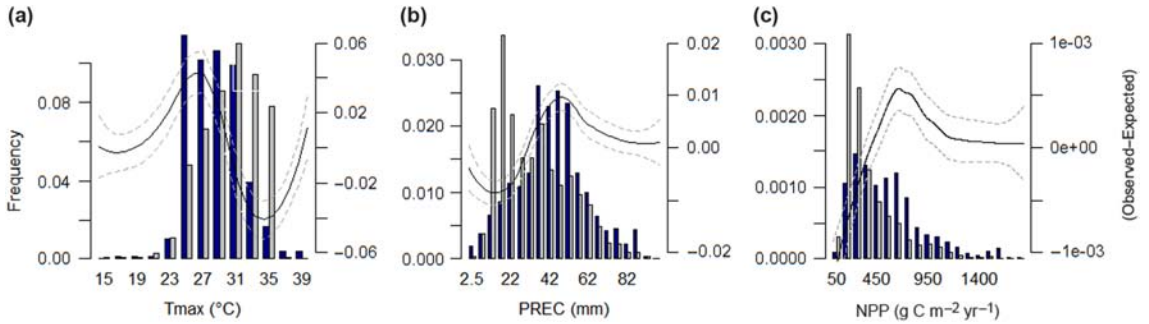

Figure 2. Frequency distribution of (a) maximum temperature of the warmest month, (b) mean monthly precipitation, (c) mean annual net primary productivity in historical written records (OBS – blue bars) and background records (BACKGROUND – grey bars). The black lines represents the difference between observed and expected frequencies, smoothed by local regression (plain line), and the 95% confidence intervals (dashed lines). The values for the smoothed regression lines are shown on the secondary y axis. A difference above (below) 0 means that observed frequency is higher (lower) than expected from a random space use. The observations cover more than 95% of the environmental range for the three environmental variables considered.

were aggregated on a 0.1° × 0.1° grid, using the raster package (Hijmans 2014) in R 3.4.3. They include information on three levels of seasonality of river flow (perennial, non-perennial, dry) that were used to adjust the model.

For each cell of the study area, we calculated the Euclidean distance to the nearest settlement and freshwater source, respectively $dist_S$ and $dist_W$. Location of rivers itself is a poor indicator of the proximity of water, as the seasonality of river flow varies in space (Uys and O'Keeffe 1997). Distance to freshwater was thus weighted by the probability of finding water in that river, respectively assigned a value of 0.01, 0.5 and 1 for dry, non-perennial and perennial rivers. For each cell x_i of the study area, we then calculated the settlement proximity index (*SPI*) and water proximity index (*WPI*) as a Gaussian kernel function of the distance to both features. The accessibility index *AI* was then calculated as the mean of *SPI* and *WPI*:

$$AI(x_i) = \frac{1}{2}\left(\underbrace{e^{\frac{-1}{2}\left(\frac{dist_S}{h_S}\right)^2}}_{SPI} + \underbrace{e^{\frac{-1}{2}\left(\frac{dist_W}{h_W}\right)^2}}_{WPI} \right)$$

Where h_S *and* h_W are the kernel widths for settlements and freshwater, respectively, summarizing the scale at which the feature influences the movement of observers.

To calibrate the model, we used an estimate of the average distance that observers could cover in one day. According to Burman (1988, in Joubert 1995), a span of oxen on trek could travel at between four and five kilometers per hour, and trekkers did not exceed 10 km per day if accompanied with sheep. When no livestock were present, 36 km could be covered in winter and 24 km in summer. Given this information, we used a rough estimate of around 20 km covered daily with an ox-wagon. We set h_S at 40 km, i.e. approximately two days of travel and h_W at 10 km, i.e. about half a day of travel. These values seemed relevant in the light of our knowledge of the constraints faced by travellers (Supplementary material Appendix 1). *AI* was scaled between zero and one, higher values representing more accessible areas. The correlation between the two components of AI (WPI and SPI) is 0.13 (Spearman's correlation statistic). See Supplementary material Appendix 5 for a sensitivity test with different values of the model's parameters.

Model's ability to predict observer's occurrence

We first evaluated the model's ability to predict observer's presences by plotting the number of observed occurrences that fall in each of *n* classes of predicted AI (n = 20). From this, we calculated Pearson's correlation coefficient (ρ) between the number of observed occurrences and predicted AI, a value of ρ = 1 indicating a total positive linear correlation between the two variables.

We then measured how much the model's prediction differ from random distribution of the observed presences across the prediction gradient using the continuous Boyce index (B_{cont}), as described in Hirzel et al. (2006). This method consists in partitioning the AI range in i classes using a 'moving window' of width W (W = 1/10 of the AI range). For each class i, it calculates the frequency of evaluation points predicted by the model to fall in this class (Pi) and the expected frequency from a random distribution across the study area (Ei). For each class, the predicted-to-expected (P/E) ratio is calculated (Boyce et al. 2002). Low AI class should contain fewer evaluation presences than expected by chance, resulting in P/E < 1 whereas high AI classes should have P/E increasingly higher than 1. For a model that properly predicts presences, the plot of P/E against AI is expected to show a monotonically increasing curve, i.e. P/E increase as AI increases. This monotonic increase is measured by the continuous Boyce index (B_{cont}), calculated as the Spearman rank correlation coefficient between P/E and AI. B_{cont} varies from –1 to 1, with 0 indicating a random model and 1 a perfect agreement between the prediction and the data.

As the distribution of settlements changed over time, so did the accessibility of the landscape. We calculated AI at different time periods (Pre-1720, 1721–1760, 1761–1800, 1801–1840, 1841–1880 and 1881–1920), taking into account the establishment date of European settlements. We then evaluated the model's ability to predict the presence of observers for each time period using the same methods as described above (see the results of this analysis in Supplementary material Appendix 3).

We also tested the predictive performance of each components of AI with two other models: the SPI model (WPI set to 0) and the WPI model (SPI set to 0) using the same set of model performance estimators (Supplementary material Appendix 5), and investigated the ability of the model to predict the density of records (Supplementary material Appendix 6).

Comparison of environmental biases in the model and the data

To be pertinent in addressing sampling bias in spatial modeling approaches, the accessibility map should share similar environmental biases as the occurrence data, for the environmental variables that may be used in the model (Phillips et al. 2009). We compared environmental biases in three different datasets: OBS, BACKGROUND and MODEL (the latter being defined as all cells in the study area, n = 8387, to which we assigned values of predicted AI as weights). We considered the three previously mentioned environmental variables (Tmax, PREC and NPP), plus the South African biomes (BIOMES), which are simplified vegetation units defined on floristic criteria, exposed to similar macroclimatic patterns, with broad scale applicability to develop conservation and management strategies over large areas (Mucina and Rutherford, 2006). We acquired spatial information on biomes from the 2012 Vegetation Map of South Africa, Lesotho and Swaziland (Mucina and Rutherford 2006, South African National Biodiversity Inst. 2012). The Savanna biome in our study area is essentially composed

of the Eastern Kalahari Bushveld and Kalahari Duneveld Bioregions, which present excessively drained sandy soil with high evaporation rates, and are thus considered Arid Savanna (Mucina and Rutherford 2006). We extracted environmental values for each location of the three datasets and plotted the estimated mean and 95% confidence interval from a linear regression, weighted with values of AI for the MODEL dataset. This allowed us to compare environmental biases in the OBSERVED and MODEL datasets, compared to what would be expected given a uniform use of the study area (BACKGROUND). We tested the difference in mean environmental values for Tmax, PREC and NPP between the three datasets using weighted two-sample t-tests (Bland and Kerry 1998). Finally, we plotted the weighted histogram of the frequency distribution of BIOMES in each dataset (the weights for the MODEL dataset being the extracted values of predicted AI for each cell of the study area).

Data deposition

The spreadsheet of historical occurrence records, the GIS layers and the R code used in this paper are available from Figshare Digital Repository (<https://doi.org/10.6084/m9.figshare.c.3916342.v1>) (Monsarrat et al. 2018)

Results

Historical occurrence records are widely but not uniformly distributed in space. They are spatially clustered compared to what would be expected from a random point process (Supplementary material Appendix 2), with an apparent bias towards southern coastal areas and the north-east of the study area (Fig. 1). Records are confined to the southern part of the country until 1760, after which a shift in distribution of records towards the north and east of the study area is observed (Supplementary material Appendix 3 Fig. A3.1). The number of records collected in 40-year periods increases from 1720 to 1840 (884 records in 1800–1840) but decreases afterwards, with only 326 records between 1880 and 1920 (Supplementary material Appendix 3 Fig. A3.1). We found environmental biases in the historical records when compared to the availability of environmental conditions in the study area (Fig. 2) (Mann-Whitney U test, $p < 0.001$), with a higher sampling of areas with low maximum temperature of the warmest month, high mean monthly precipitation and high mean annual net primary productivity. The observations cover more than 95% of the environmental range for the three environmental variables considered (Tmax, PREC, NPP).

The model predicts areas of high accessibility in the southern and eastern parts of the study area. On the contrary, the northern part of the current Northern Cape and the western part of the Northwest Province show low accessibility indexes (Fig. 3). There is a strong linear correlation between the frequency of observed occurrences and predicted AI (Fig. 4a; Pearson's correlation coefficient, $\rho = 0.93$). The plot of P/E against AI (Fig. 4b) and the high continuous Boyce index value ($B_{count} = 0.995$) show an extremely high ability of the model to predict observer's presences. A similar analysis based on the density of records per cell was much less predictive, due to the large amount of cells with few records (Supplementary material Appendix 6). When evaluating the predictive ability of the model at different time periods, we found that the model performed better in the very early phase of South Africa colonization (before 1720) and in the recent past (after 1840) than in the 1720–1840 period, which corresponds to the exploration phase of South Africa (i.e. when European settlements began to be built but were not yet officially established) (Supplementary material Appendix 3).

The comparison of environmental biases in the BACKGROUND, OBS and MODEL datasets shows that both OBS and MODEL have different estimates of mean environmental values than would be expected from a uniform use of the study area (BACKGROUND) (weighted paired t-test, $p < 0.001$) (Fig. 5a–c). The estimates of mean environmental values of OBS and MODEL are also different but they are consistently more similar between each other than between OBS and BACKGROUND. Biomes are not sampled evenly, with arid environments like Nama-Karoo, Succulent Karoo and Savanna being less represented in the dataset than expected by a uniform use of the area (Fig. 5d), whereas the Albany Thicket, Azonal Vegetation, Fynbos and Grassland are positively selected. For all biomes, the distribution of frequencies of OBS and MODEL are consistently biased in the same direction when compared to BACKGROUND (Fig. 5d).

Discussion

Sampling bias can be difficult to quantify in small, single-taxa datasets of species occurrence. We show that an accessibility map built with a model based on very simple statistical rules and only two spatial features can approximate the spatial distribution and environmental biases observed in an empirical dataset of historical written occurrences. These results suggest that sampling effort can be modelled accurately without the use of empirical data, given that we know the processes influencing the bias behind data collection. This has strong implications for the inclusion of small historical datasets in ecological and conservation studies.

Assumptions and limitations

Our assumption that the distribution of historical written records of large mammal occurrence is representative of the true distribution of observers in the early historical period could be challenged if this distribution reflects a biological reality rather than sampling effort. For example, some habitats in South Africa could have no large mammal species (hence the absence of records would represent an absence of mammal species rather than of observers) or the bias towards the proximity of freshwater could be caused by a higher

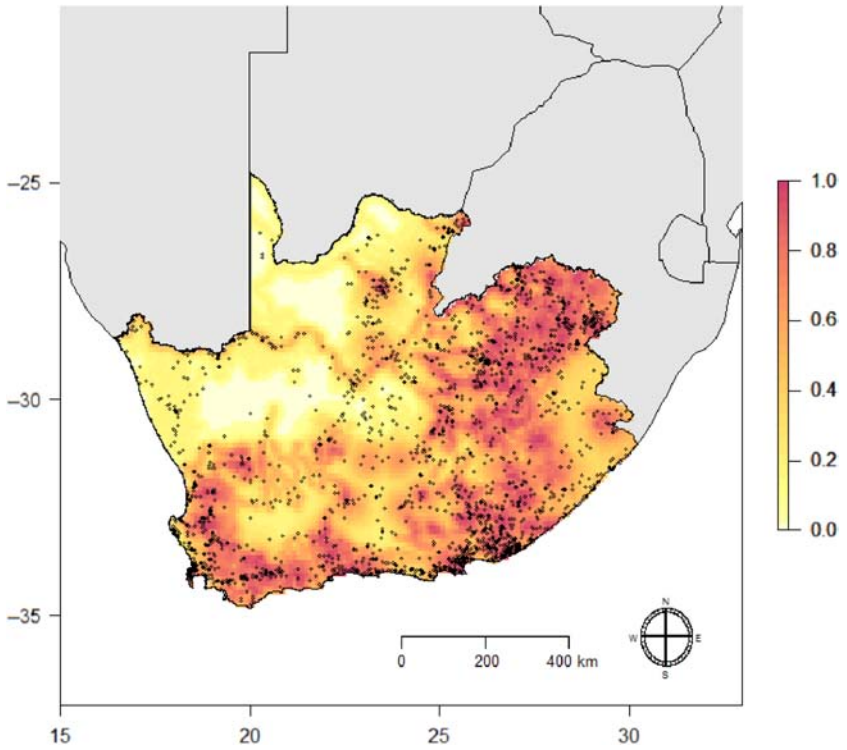

Figure 3. Accessibility map and historical data. The map of accessibility index (AI) for observers in the early historical period (1497–1920) in the study area is built from a model based on two spatial features, proximity of freshwater and proximity of European settlements. Shades of red indicate progressively higher accessibility as predicted by the model. Black dots are historical written records of large mammal occurrence for the period 1497–1920.

detectability of mammal species around water resources. However, the 51 mammal species included in the dataset are distributed throughout South Africa, occupy a great variety of habitats and biomes and have very different ecologies (Skinner and Chimimba 2005). For example, some species that are well adapted to arid environments (e.g. springbok, brown hyaena, eland, gemsbok) (Skinner et al. 1984, Hofmeyr and Louw 1987, Skinner and Chimimba 2005)

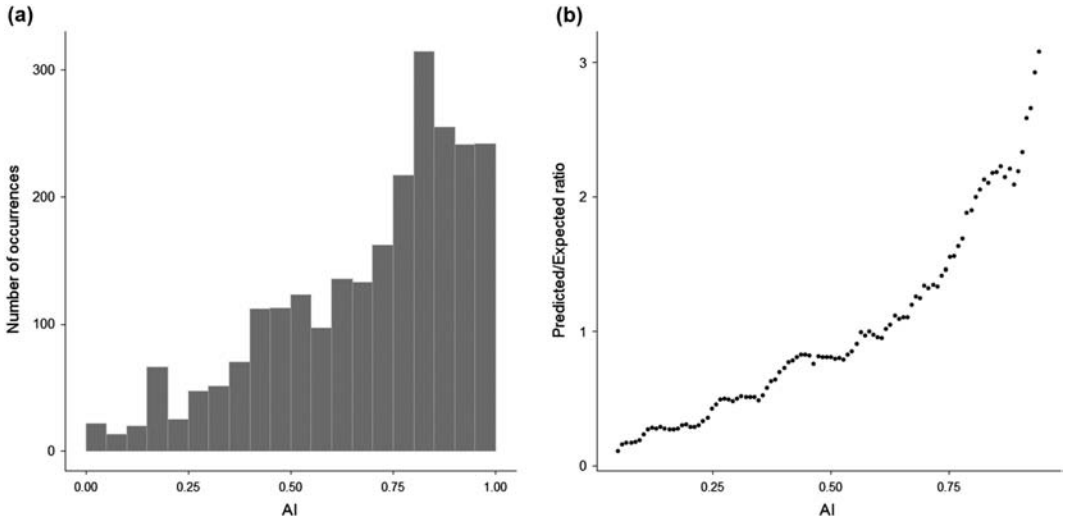

Figure 4. Model performance plots. (a) Histogram of number of occurrences against the predicted accessibility index (AI). The value of Pearson's coefficient of correlation between observed frequency of presences and predicted AI is 0.93. (b) Predicted/Expected (P/E) curve. Each point is calculated as the ratio of frequency of evaluation points predicted by the model (P) and the expected frequency from a random distribution across the study area (E) for the corresponding AI class. A straight curve indicates an ideal model with perfect predictive ability. The measure of monotonic increase of the P/E curve calculated with the Spearman's correlation coefficient gives a continuous Boyce index B_{cont} of 0.995.

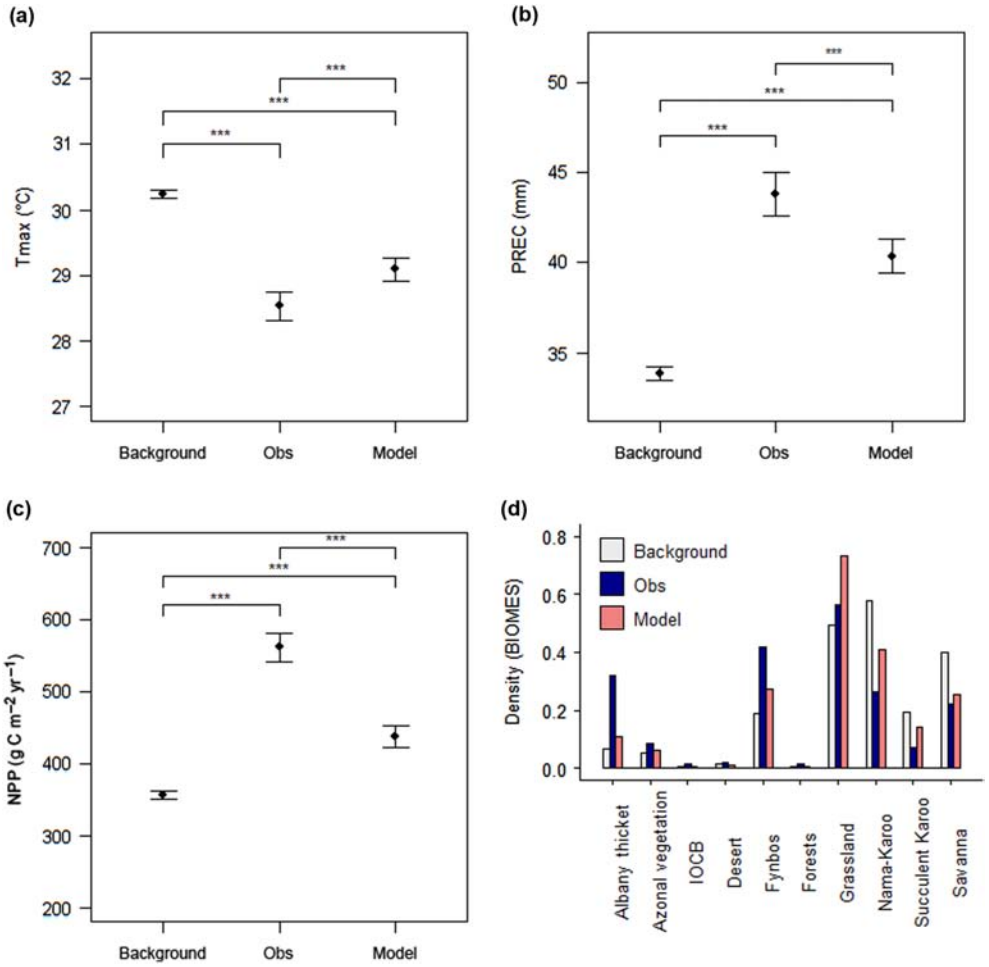

Figure 5. Comparison of environmental biases in the 'background', 'obs' and 'model' datasets. (a) to (c) show the modelled estimate of mean values and 95% confidence interval for maximum temperature of the warmest month (Tmax), mean monthly precipitation (PREC) and mean annual net primary productivity (NPP), respectively. Levels of significance for the differences between the three groups from weighted paired t-tests are indicated in each graph (***: $p<0.001$); (d) shows the frequency distribution of biomes in the 'background' (grey), 'obs' (blue) and 'model' (pink) datasets (IOCB: Indian ocean costal belt).

occur naturally in areas that are predicted to have limited accessibility in the model. These species are generally over-reported in historical accounts (Monsarrat and Kerley 2018) and they are reported elsewhere in the study area, indicating that the absence of records in arid areas is due to a lack of effort rather than to the absence of mammal species. Also, some species are water-independent and their detectability would not necessarily increase close to freshwater. However, there remains the possibility of a confusion between species' detectability and bias in sampling effort. While this limitation is difficult to test, it should be kept in mind in the interpretation of the results.

Ecological responses to climate change over the past centuries (Parmesan 2006, Wanner et al. 2008) may affect the results of our analyses, in a way that we are unable to account for. We used a long-term climatology to mitigate the impact of inter-annual variations in climatic variables. Additionally, because all values in the different datasets have been extracted from the same environmental data, the relative comparison between the OBS, MODEL and BACKGROUND datasets remains valid.

We found a good but not perfect agreement between the predicted map of accessibility and the location of observers' records. Some records located in areas with a very low predicted accessibility suggest that additional factors influence the distribution of observers. It is likely that the establishment of roads in the early 20th century would have influenced the accessibility of the area, and hence the location of records, at that period, as suggested by studies on modern correlates of sampling bias (Kadmon et al. 2004). However, in the absence of reliable data on the network of roads pre-1920, we were unable to test this idea. Using the modern road network as a proxy would be problematic as many modern cities were not yet established in 1920 and the road network would have changed considerably in the 20th century. One of the main factors that we could not include in the model are the travelers' motivation. For example, in 1795, the French naturalist François Le Vaillant travelled

from Cape Town to the Orange River, crossing the arid Karoo plains of the Western and Northern Cape in order to be the first European to describe the fauna in this remote area. The author describes how he was obliged to abandon his three wagons and leave his people and baggage dispersed on the road, after most of his 52 oxen died and he and his company almost died of thirst (Le Vaillant 1796). While it cannot predict such idiosyncratic behavior, the model shows nonetheless a high ability to correctly predict positive occurrences in the study area, suggesting that it can be a powerful tool to assist in the interpretation of historical observers' records.

Biases in South African historical written records of mammal occurrence

We provide evidence of spatial, temporal and environmental biases in written records of mammal occurrence collected in the early historical period in South Africa. Travelers' avoidance of hot, arid and unproductive habitats is an intuitive result, as they would require large amount of water and fodder for their livestock. The changes in distribution of observer effort over time can be explained by the colonial history of South Africa (Supplementary material Appendix 3). The spatio-temporal distribution of these historical occurrence records raises questions about how our baseline of the historical composition of mammal communities is shifted spatially – in addition to temporally – as different areas are sampled in different time periods. When using historical written records to reconstruct ecological baselines, one has to bear in mind the spatio-temporal context of the observers' distribution. For example in South Africa, the interior part of the country was never described in writing before 1760, and the Orange Free State was mostly unknown to literate travellers before the beginning of the 19th century. Other sources (e.g. archaeological, palaeontological) should therefore be considered to study earlier ecological conditions in these regions.

Knowing about temporal, geographic and environmental biases is important to improve the use of these historical records in South Africa conservation research and management. One can use this information to identify gaps in knowledge to guide future data collection effort (e.g. looking for archaeological records in areas with a low density of records to inform the historical distribution of these species) or to develop appropriate analytical tools for these data (e.g. by explicitly addressing biases in spatial analyses). This provides opportunities to go beyond descriptive approaches (e.g. as in Boshoff et al., 2016) and to develop tools to derive historical species distribution from these written historical records. Additionally, these historical sources contain descriptions on the fauna, vegetation, climate and inhabitants of the country (Rookmaaker 1989, Hoffman et al. 1995, Nash and Endfield 2002, Huigen 2009, Skead 2009). Thus, the study of sampling bias can be informative across disciplines for data extracted from these sources.

Relevance of accessibility maps to quantify sampling biases

The analysis of historical occurrence records with spatial modeling methods has great potential to improve our knowledge of species' ecology, historical distribution and status changes, with implications for their conservation and the management of remaining populations. However, if used without correction for sampling bias, species status changes or conservation plans based on such uncorrected data may be misleading.

Accessibility maps can be useful tools to report the spatial distribution of bias and lack of sampling effort across a study region, providing an innovative avenue to improve the statistical power of spatial analyses based on small datasets of species occurrences. In contrast with other approaches used to estimate sampling bias, accessibility maps require no empirical data of observer occurrence. Their applications are similar to those of uncertainty maps (Rocchini et al. 2011, Ruete 2015) or maps of sampling effort (Stolar and Nielsen 2015). These include the visual exploration of the quality of the data and the improvement of inferences made from them (Tingley and Beissinger 2009). Following existing methods to address sampling bias in species distribution models (Phillips et al. 2009, Hertzog et al. 2014), accessibility maps could be used to manipulate background data in species distribution modelling to generate pseudo-absences data with a similar geographical sampling bias to that of the presence data. It could also be used to adjust model estimates by down-weighting sample points from locations with higher accessibility (Stolar and Nielsen 2015).

Here, we show that proximity of freshwater and proximity of settlements alone can be good predictors of sampling bias in historical occurrence data. This is both a surprising outcome – meaning that sampling effort can be predicted with very simple rules and that understanding biases in other contexts may not require 'rocket science' – and an expected result, as it is consistent with previous empirical analyses of sampling biases in occurrence data (Hijmans et al. 2000, Soberón et al. 2000, Reddy and Davalos 2003, Newbold 2010). We believe that the impact of freshwater and settlement proximity on observers' movements is not specific to the South African context and that these environmental features are also limiting for early travelers in other parts of the world. Freshwater availability would be particularly limiting in arid environments, but the potential for rivers to serve as means of transportation is not to be underestimated in other, less dry, areas. The importance of each factor, however, can vary over time. During the exploration phase of South Africa, when settlements were already in place but not officially established, freshwater alone seemed a better predictor of traveler's distribution. This is due to a lag between the building of European settlements and their official date of establishment. For a given time period, it is thus important to consider localities that were important aggregation places, in addition to officially established settlements, in order to draw a better picture of the availability of the area at that time. Understanding the constraints that historical observers faced

in their travel might require reading historical references that describe the lifestyle and habits of early travelers. A huge amount of historical references that are digitized and available online through archives (e.g. Internet Archive <https://archive.org/>, Google Books <https://books.google.com>, the Biodiversity Heritage Library <www.biodiversitylibrary.org>) could be used for that purpose.

While this study focuses on historical written records of occurrence, a parallel can be drawn with contemporary datasets that are subject to the same type of biases. Amateur naturalists still collect opportunistic sightings, in what is now defined as citizen science (Dickinson et al. 2010). Historical occurrence records can therefore be seen as a particular case of citizen-derived science data (Miller-Rushing et al. 2012), both types of data presenting similar sampling biases. Additionally, modern datasets of occurrence for some under-represented taxa such as insects, and invertebrates in general, may share similar limitations with historical datasets on vertebrates (Lobo and Martin-Piera 2002). The approach described in this study could thus be added to the existing toolbox of methods used to address biases in citizen science datasets (Bird et al. 2014, Isaac et al. 2014) or other small datasets of occurrence, making this analysis relevant to a much broader audience than the community of historical ecologists.

We encourage further testing of this approach in different spatio-temporal contexts and for other taxa to 1) evaluate the ability of accessibility maps to predict sampling bias in other large datasets of occurrence and 2) compare the performance of species distribution models using accessibility maps vs other existing methods to correct for sampling bias. If accessibility maps prove to be robust predictors of sampling bias in different contexts and are shown to improve the performance of species distribution models, this will provide strong support for their relevance in addressing sampling bias in the analyses of small datasets of occurrence.

Specific recommendations

We suggest that, rather than discarding small historical occurrence datasets a priori due to possible biases they may contain, researchers and conservationists could use accessibility maps to explore sampling bias and improve the use of these data in modern quantitative analyses. The strength of this tool lies in its simplicity, and on its non-reliance on empirical data of observer occurrence. The only requirements to build accessibility maps are to have 1) a good knowledge of the processes underlying the behavior and environmental constraints faced by observers, which can be inferred from the historical literature and 2) access to spatially explicit information on the distribution of relevant environmental features. From this, spatially explicit functions describing the accessibility of the study area can be used to calculate the accessibility index for the study area.

This approach can easily be implemented and applied to other spatio-temporal contexts. We recommend that users thoroughly identify which environmental features constrain the movement of observers in order to set appropriate rules to build the accessibility map. Proximity to freshwater and to settlements seem important features in South Africa in the 15th–19th centuries, but other parameters may influence the distribution of observers in other contexts. For example, avoidance of diseases and social conflicts, natural barriers, attraction for minerals or natural resources are examples of other elements that could be included in the model, depending on the context of the study. The temporal and spatial extent and resolution of the analysis should also be adapted so as to produce accessibility maps that are relevant to the ecology of the study model and researchers' needs.

Acknowledgements – We would like to thank Krista Oswald and Hugo Valls for useful comments on earlier versions of the manuscript.
Funding – This work was supported by a post-doctoral fellowship from the Claude Leon Foundation.

References

Andrews, P. and O'Brien, E. M. 2000. Climate, vegetation, and predictable gradients in mammal species richness in southern Africa. – J. Zool. 251: 205–231.

Baddeley, A. et al. 2014. On tests of spatial pattern based on simulation envelopes. – Ecol. Monogr. 84: 477–489.

Bird, T. J. et al. 2014. Statistical solutions for error and bias in global citizen science datasets. – Biol. Conserv. 173: 144–154.

Bland, J. M. and Kerry, S. M. 1998. Weighted comparison of means. – BMJ 316: 129.

Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. – Ecol. Model. 275: 73–77.

Boshoff, A. F. and Kerley, G. I. 2010. Historical mammal distribution data: how reliable are written records? – South Afr. J. Sci. 106: 26–33.

Boshoff, A. F. and Kerley, G. I. H. 2013. Historical incidence of the larger mammals in the Free State Province (South Africa) and Lesotho. – Centre for African Conservation Ecology, Nelson Mandela Metropolitan Univ.

Boshoff, A. F. and Kerley, G. I. H. 2015. Lost herds of the Highveld: evidence from the written, historical record. – Afr. J. Wildl. Res. 45: 287–300.

Boshoff, A. F. et al. 2002. The potential distributions, and estimated spatial requirements and population sizes, of the medium to large-sized mammals in the planning domain of the Greater Addo Elephant National Park project. – Koedoe 45: 85–116.

Boshoff, A. et al. 2016. Filling the gaps on the maps: historical distribution patterns of some larger mammals in part of southern Africa. – Trans. R. Soc. South Afr. 71: 1–65.

Boyce, M. S. et al. 2002. Evaluating resource selection functions. – Ecol. Model. 157: 281–300.

Burman, J. 1988. Towards the far horizon: the story of the ox-wagon in South Africa. – Human and Rousseau.

Butynski, T. M. et al. 2015. Historic and current distribution, abundance, and habitats of Roosevelt's sable antelope *Hippotragus niger roosevelti* (Heller, 1910) (*Cerartiodactyla: Bovidae*) in Kenya. – J. East Afr. Nat. Hist. 104: 41–77.

Clavero, M. and Delibes, M. 2013. Using historical accounts to set conservation baselines: the case of Lynx species in Spain. – Biodivers. Conserv. 22: 1691–1702.
Clavero, M. and Revilla, E. 2014. Biodiversity data: mine centuries-old citizen science. – Nature 510: 35–35.
Coe, M. J. et al. 1976. Biomass and production of large African herbivores in relation to rainfall and primary production. – Oecologia 22: 341–354.
Dickinson, J. L. et al. 2010. Citizen science as an ecological research tool: challenges and benefits. – Annu. Rev. Ecol. Evol. Syst. 41: 149–172.
Dixon, P. M. 2002. Ripley's K function. – Encyclop. Environmetrics 3: 1796–1803.
Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.
Floyd, T. B. 1960. Town planning in South Africa. – Shuter and Shooter.
Fourcade, Y. et al. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. – PLoS One 9: e97122.
Hertzog, L. R. et al. 2014. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. – Divers. Distrib. 20: 1403–1413.
Hijmans, R. J. 2014. raster: geographic data analysis and modeling. – <https://cran.r-project.org/web/packages/raster/index.html>.
Hijmans, R. et al. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. – Conserv. Biol. 14: 1755–1765.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hirzel, A. H. et al. 2006. Evaluating the ability of habitat suitability models to predict species presences. – Ecol. Model. 199: 142–152.
Hoffman, M. T. et al. 1995. Desertification of the eastern Karoo, South Africa: conflicting paleoecological, historical, and soil isotopic evidence. – Environ. Monit. Assess. 37: 159–177.
Hofmeyr, M. and Louw, G. N. 1987. Thermoregulation, pelage conductance and renal function in the desert-adapted springbok, *Antidorcas marsupialis*. – J. Arid Environ. 13: 137–151.
Hoving, C. L. et al. 2003. Recent and historical distributions of Canada Lynx in Maine and the Northeast. – Northeast. Nat. 10: 363.
Huigen, S. 2009. Knowledge and colonialism: eighteenth-century travellers in South Africa. – Brill.
Isaac, N. J. B. et al. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. – Methods Ecol. Evol. 5: 1052–1060.
Joubert, B. 1995. An historical perspective on animal power use in South Africa. – In: Starkey, P. (ed.), Animal power in South Africa: empowering rural communities. Development Bank of Southern Africa, Gauteng, South Africa, pp. 125–138.
Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Kittinger, J. N. et al. 2013. Using historical data to assess the biogeography of population recovery. – Ecography 36: 868–872.
Le Vaillant, F. 1796. New travels into the interior parts of Africa: by the way of the Cape of Good Hope, in the years 1783, 84 and 85. – G. G. and J. Robinson, London.
Lobo, J. M. and Martin-Piera, F. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. – Conserv. Biol. 16: 158–173.
Matthews, P. E. and Heath, A. C. 2008. Evaluating historical evidence for occurrence of mountain goats in Oregon. – Northwest Sci. 82: 286–298.
Mihoub, J.-B. et al. 2017. Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures. – Sci. Rep. 7: 41591.
Miller-Rushing, A. et al. 2012. The history of public participation in ecological research. – Front. Ecol. Environ. 10: 285–290.
Monsarrat, S. et al. 2018. Data from: Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. – Figshare Digital Repository, <https://doi.org/10.6084/m9.figshare.c.3916342.v1>.
Monsarrat, S. and Kerley, G. I. H. (2018) Charismatic species of the past: biases in reporting of large mammals in historical written sources. – Biol. Conserv. 223: 68–75.
Mucina, L. and Rutherford, M. C. 2006. The vegetation of South Africa, Lesotho and Swaziland, Strelitzia. – SANBI, Pretoria.
Nash, D. J. and Endfield, G. H. 2002. A 19th century climate chronology for the Kalahari region of central southern Africa derived from missionary correspondence. – Int. J. Climatol. 22: 821–841.
Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. – Prog. Phys. Geogr. 34: 3–22.
Parmesan, C. 2006. Ecological and evolutionary responses to recent climate change. – Annu. Rev. Ecol. Evol. Syst. 37: 637–669.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Ranc, N. et al. 2016. Performance tradeoffs in target-group bias correction for species distribution models. – Ecography 40: 1076–1087.
Reddy, S. and Davalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Rick, T. C. and Lockwood, R. 2013. Integrating paleobiology, archeology, and history to inform biological conservation. – Conserv. Biol. 27: 45–54.
Rocchini, D. et al. 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. – Prog. Phys. Geogr. 35: 211–226.
Rookmaaker, L. C. 1989. The zoological exploration of Southern Africa 1650–1790. – Taylor and Francis.
Rookmaaker, L. C. 2007. A chronological survey of bibliographical and iconographical sources on rhinoceroses in southern Africa from 1795 to 1875: reconstructing views on classification and changes in distribution. – Trans. R. Soc. South Afr. 62: 55–198.
Ruete, A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. – Biodivers. Data J. 3: e5361.
Scholte, P. 2012. Using the past to manage for the future: contributions of early travel literature to African historical ecology. – Afr. J. Ecol. 50: 117.
Schulman, L. et al. 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation:

Amazonian collecting and range estimation. – J. Biogeogr. 34: 1388–1399.
Shaffer, H. B. et al. 1998. The role of natural history collections in documenting species declines. – Trends Ecol. Evol. 13: 27–30.
Skead, C. J. 1980. Historical mammal incidence in the Cape Province, vol. 1: the western and northern Cape. – Dept of Nature and Environmental Conservation of the Provincial Administration of the Cape of Good Hope, Cape Town, South Africa.
Skead, C. J. 1987. Historical mammal incidence in the Cape province, vol. 2: The eastern half of the Cape. – Dept of Nature and Environmental Conservation of the Provincial Administration of the Cape of Good Hope, Cape Town, South Africa.
Skead, C. J. 2009. Historical plant incidence in southern Africa: a collection of early travel records in southern Africa. – South African Natl Biodiversity Inst.
Skead, C. J. et al. 2007. Historical incidence of the larger land mammals in the broader eastern Cape, 2nd ed. – Centre for African Conservation Ecology, Nelson Mandela Metropolitan Univ., Port Elizabeth.
Skead, C. J. et al. 2011. Historical incidence of the larger land mammals in the broader western and northern Cape. – Centre for African Conservation Ecology, Nelson Mandela Metropolitan Univ., Port Elizabeth.
Skinner, J. D. and Chimimba, C. T. 2005. The mammals of the southern African sub-region. – Cambridge Univ. Press.
Skinner, J. D. et al. 1984. Adaptations in three species of large mammals (*Antidorcas marsupialis, Hystrix africaeaustralis, Hyaena brunnea)* to arid environments. – South Afr. J. Zool. 19: 82–86.
Soberón, J. M. et al. 2000. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. – Biodivers. Conserv. 9: 1441–1466.
South African National Biodiversity Inst. 2012. 2012 Vegetation Map of South Africa, Lesotho and Swaziland [vector geospatial dataset]. – <www.sanbi.org/biodiversity/foundations/national-vegetation-map/>.
Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.
Syfert, M. M. et al. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. – PLoS One 8: e55158.
Szabó, P. 2015. Historical ecology: past, present and future. – Biol. Rev. 90: 997–1014.
Szabó, P. and Hédl, R. 2011. Advancing the integration of history and ecology for conservation: history, ecology, and conservation. – Conserv. Biol. 25: 680–687.
Tingley, M. W. and Beissinger, S. R. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. – Trends Ecol. Evol. 24: 625–633.
Turvey, S. T. et al. 2015. Historical data as a baseline for conservation: reconstructing long-term faunal extinction dynamics in Late Imperial–modern China. – Proc. R. Soc. B 282: 20151299.
Turvey, S. T. et al. 2017. Long-term archives reveal shifting extinction selectivity in China's postglacial mammal fauna. – Proc. R. Soc. B 284: 20171979.
Uys, M. C. and O'Keeffe, J. H. 1997. Simple words and fuzzy zones: early directions for temporary river research in South Africa. – Environ. Manage. 21: 517–531.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.
Wanner, H. et al. 2008. Mid- to Late Holocene climate change: an overview. – Quat. Sci. Rev. 27: 1791–1828.
Weepener, H. L. et al. 2012. The development of a hydrologically improved digital elevation model and derived products for South Africa based on the SRTM DEM. – Water Res. Comm. Pretoria South Afr. Tech. Rep. 1908/1/11: <www.dwaf.gov.za/iwqs/gis_data/river/rivs500k.aspx>.
Williams, P. H. et al. 2002. Data requirements and data sources for biodiversity priority area selection. – J. Biosci. 27: 327–338.
Willis, K. J. et al. 2007. How can a knowledge of the past help to conserve the future? Biodiversity conservation and the relevance of long-term ecological studies. – Phil. Trans. R. Soc. B 362: 175–187.
Zhao, M. et al. 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. – Remote Sens. Environ. 95: 164–176.

Supplementary material (Appendix ECOG-03944 at <www.ecography.org/appendix/ecog-03944>). Appendix 1–6.