



Charismatic species of the past: Biases in reporting of large mammals in historical written sources



Sophie Monsarrat*, Graham I.H. Kerley

Centre for African Conservation Ecology, Nelson Mandela University, PO Box 77000, Port Elizabeth, 6031, South Africa

ARTICLE INFO

Keywords:

Taxonomic bias
Charisma
Historical records
South Africa
Terrestrial mammals

ABSTRACT

Long-term biodiversity occurrence records are key to quantify long-term biodiversity patterns and trends and inform the conservation of threatened species, but they are strongly biased in terms of the species represented. This taxonomic bias, and its correlation to societal preferences, is well-identified in modern biodiversity datasets. However, it remains to be investigated, and its basis understood, in long-term occurrence datasets assembled from historical sources. Here we investigate taxonomic bias for 38 species of large terrestrial mammals using a dataset of 780 historical occurrence assembled from 16th to mid-19th century historical written sources in South Africa. We test if this bias is related to species' historical charisma, using a functional definition of non-human charisma, supported by anecdotes from the historical literature. We identify a strong taxonomic bias, with up to several order of magnitudes of difference in the likelihood of reporting between some species. Species' charisma alone explains 75% of the observed variance, the most charismatic species being largely over-reported. This is the first evidence of a positive relationship between taxonomic bias and charisma in a historical biodiversity dataset, within a homogeneous taxonomic group such as large terrestrial mammals. These results improve our understanding of the relationship between people and the large terrestrial fauna in historical times and suggest that species' charisma is a good predictor of taxonomic bias in long-term biodiversity datasets. This provides background for modern conservation by illustrating the durability of the charisma concept and of its relation with taxonomic bias, with implications for the representativeness of species in long-term conservation studies.

1. Introduction

Historical biodiversity datasets are key to detect and quantify long-term human impacts on biodiversity and inform the conservation of threatened species (Willis et al., 2007; Tingley and Beissinger, 2009; Turvey et al., 2015; Mihoub et al., 2017). Historical species lists, and particularly variations in the number of species counted at a site over time, may reflect biologically meaningful patterns of past communities of earlier ecosystems, which can be used to investigate species declines (Szabo et al., 2010). However, recorded differences in species richness may also reflect sampling biases that naturally arise from data collected opportunistically without modern sampling protocols. Of the four types of biases identified in long-term biodiversity datasets – geographical, environmental, temporal and taxonomic (Soberón et al., 2000; Newbold, 2010) – the latter has been the least investigated. However, taxonomic bias can lead to strong misconceptions of what communities and ecosystems used to look like, such gaps in knowledge affecting our understanding of biodiversity patterns and response to changes (McKinney, 1999; Feeley et al., 2017).

Taxonomic bias in conservation science has long been recognized (Clark and May, 2002) but the underlying processes that cause it are unclear. Previous studies that have investigated correlates of taxonomic bias in biodiversity datasets show a bias towards species that are more locally abundant (Royle and Nichols, 2003), or easily identified (Boakes et al., 2016). A recent study identified a strong correlation between societal preferences and taxonomic bias (Troudet et al., 2017), with the most popular species being also the species with the most records in biodiversity databases. These studies typically compared higher taxonomic groups (e.g. mammals, birds, reptiles, insects, plants, invertebrates) in modern datasets. The evolution of taxonomic bias over time has rarely been examined (Troudet et al., 2017) and, to our knowledge, no study has investigated if its relation to societal preferences is stable through time.

In southern Africa, historical accounts written by European settlers, missionaries, naturalists and explorers of the 16th to 20th century provide valuable information on the past composition of mammal fauna, but the taxonomic biases in these records remain to be investigated and understood. To report a species' presence in written

* Corresponding author.

E-mail addresses: sophiemonsarrat@gmail.com (S. Monsarrat), graham.kerley@mandela.ac.za (G.I.H. Kerley).

accounts, an observer has to 1) detect the species, whether through direct sightings or indirect cues (e.g. footprints, scats, sound, presence of burrows), and 2) be willing to report it. The latter condition is probably driven by his or her own perception of the different species he or she encounters in the field. Similar to modern societal preferences (Troudet et al., 2017), species' charisma might be a key trait that contributes to taxonomic bias in historical reporting of biodiversity.

Wildlife charisma is a modern concept with strong implications for the use of species as flagships for conservation (Leader-Williams and Dublin, 2000), as attractions for tourism (Goodwin and Leader-Williams, 2000), and even as a marketing strategy (Feldhamer et al., 2002). The term *charismatic* was first coined for nonhuman use in the conservation literature in the 1980's (Myers, 1983; Western, 1987), and has since been used as one of the traits to identify flagship species in conservation (Heywood, 1995; but see Verissimo et al., 2011). While many recent articles use this expression, there is little consensus on a functional definition (Ducarme et al., 2013). In an attempt to map nonhuman charisma, Lorimer (2007) defined it as “the distinguishing properties of a nonhuman entity or process that determine its perception by humans and its subsequent evaluation”, insisting that the perception of charisma is subjective and dependent on the human population considered. Lorimer (2006, 2007) then provides a typology of the factors that define nonhuman charisma, describing it as an amalgam of “detectability and distinctiveness” (DETEC - how likely people are to see or hear a species and their ability to distinguish it from similar taxa), “socioeconomic biases” (ECON - the economic costs and benefits of species to different land users), “aesthetics” (AESTH - the distinguishing properties of an organism's behaviour and appearance that trigger particular emotions, both positive or negative, in those humans it encounters), and “intellectual satisfaction” (INTELL - the emotions experienced by humans in their practical interactions with an organism or group of organisms in the field). Understanding the persistence or durability of the *nonhuman charisma* notion can help us interpret the long-term relationships of humans with nature and how present conservation values came about. As Rangarajan puts it in a review of the conservation dilemmas in Africa, “by knowing better what choices were made in the past, when and why, the dilemmas of the present can be seen in a more holistic way” (Rangarajan, 2003:77).

Here we aim to identify the taxonomic bias in a dataset of historical occurrence records of large terrestrial mammals collected from written sources of the 16th to mid-19th century in South Africa. We hypothesize that this taxonomic bias is related to the perception and attitude of people towards the large mammal fauna, approximated by species' charismatic value. We test whether this hypothesis is supported by a positive relationship between the reporting bias in the South African dataset and the perceived charisma of species.

2. Study area

We focused on the Cape Floristic Region planning domain in southwestern South Africa, an area for which we have data on the estimated historical relative abundance of each species in the dataset (Kerley et al., 2003b). It includes the Cape Floristic Region (CFR) and an extension of approximately 60 km beyond the boundaries of the CFR, for a total of 122,590 km², as described in Kerley et al. (2003b). The CFR is a global priority for conservation action and is listed as a biodiversity hotspot of global significance (Myers et al., 2000), for its diversity of endemic flora and fauna.

3. Methods

3.1. Historical occurrence records

We used a dataset of 780 historical occurrence records of large terrestrial mammals (31 species in 10 families) within the CFR planning domain, this assembled from letters, journals, diaries or books written

by literate pioneers in southern Africa (Boshoff et al., 2016). This dataset comes from a compilation of written distribution records extracted from a range of key references (e.g. Skead, 1980, 2011; Rookmaaker, 1989, 2007) and complemented with additional occurrence records previously overlooked in literature sources (Boshoff et al., 2016). The first record dates back to 1497, when the Portuguese explorer Vasco de Gama reported the presence of elephants in Mossel Bay (in Colvin, 1912). We only considered records collected before 1850, after which the impact of European settlers on the large mammal fauna in the study area increased significantly, from direct hunting pressure, increased predator control (Skead, 2011:426), loss of habitat due to the growth of the stock industry (Skead, 2011:436) and the development of roads and railways (Van Sittert, 2005:277). The reliability of these records in terms of identification and locality is discussed in Boshoff and Kerley (2010).

3.2. Taxonomic bias

We quantified the taxonomic bias in the historical dataset as the ratio between the observed relative frequency of species in the historical dataset, and their expected relative frequency based on estimated historical relative abundances, obtained from an independent study (Kerley et al., 2003b).

3.2.1. Observed relative frequency

Each occurrence in the dataset may correspond to one or more individuals observed, particularly for gregarious species. To calculate observed reported abundance, we multiplied the number of times each species appeared in the dataset by an estimate of the average group size for that species, as an estimate of the actual number of individuals seen by observers. Values of mean average group size for each species were extracted from the literature and are detailed in Appendix (Table A.1). We calculated species' observed relative frequency as the ratio between each species observed reported abundance and the sum of all species' reported abundances.

3.2.2. Expected relative frequency

Kerley et al. (2003b) estimated the potential historical abundance of the large and medium-sized mammals in the Cape Floristic Region, calculated from estimates of historical distribution and the densities, social structure, territory sizes and home ranges for carnivore species, and forage availability estimates and metabolic requirements for herbivores. Their study includes 41 large and medium-sized mammal species (mass > 2 kg) indigenous to the CFR, but excluded the hippopotamus *Hippopotamus amphibius* and the African clawless otter *Aonyx capensis* (Kerley et al., 2003b). We did not include four species that have marginal habitat in the study area (cheetah *Acinonyx jubatus*, gemsbok *Oryx gazella*, oribi *Ourebia ourebi* and warthog *Phacochoerus aethiopicus*) and extracted potential historical abundance for the remaining 37 species. We calculated each species' historical expected relative frequency as the ratio between that species' historical abundance and the sum of all species' abundances.

3.2.3. Reporting bias index

The reporting bias index (B_i) for each species (i) is the ratio between that species' observed relative frequency in the historical dataset and its expected historical relative frequency. A value of $B_i > 1$ means that the species was over reported compared to what would be expected given its estimated historical abundance (and vice versa for $B_i < 1$). Seven species were not recorded in the historical dataset despite being historically present in the study area (aardvark *Orycteropus afer*, aardwolf *Proteles cristata*, African wild cat *Felis silvestris lybica*, Cape fox *Vulpes chama*, honey badger *Mellivora capensis*, mountain reedbuck *Redunca fulvorufula* and small spotted cat *Felis nigripes*). We assigned a value of $B_i = 0.01$ to these species, representing a very low (about 1/4 of the lowest value for reported species) but non-null reporting bias index,

Table 1
Definition of the four components of species' charisma, adapted from Lorimer (2007).

Criteria	Definition
Detectability	The physical and behavioural properties that affect how observers are likely to see or hear a species and their ability to distinguish it from similar others. Relevant physical characteristics are body size, color, shape. Behavioural properties are diurnal/nocturnal activity, vocalization and habitat (e.g. species restricted to largely inaccessible habitat (e.g. mountain tops) are considered poorly detectable).
Economic bias	The economic importance of species for observers, both in terms of gains and costs. Species can be valuable for their meat, their skin and fur and their value on the international trade market (e.g. ivory, rhinoceros horns). Costs include personal risk, predation on livestock and damages to crops.
Aesthetics	The physical properties that triggered particular emotions in observers. Aesthetic charisma need not necessarily be sympathetic. Instead, it exists on a cuddly (positive) – feral (negative) continuum, where banal species in the middle will trigger little emotion and be given a low aesthetic score.
Intellectual satisfaction	The impression and emotions – positive (awe) or negative (fear, hate) – that an animal caused through their practical interactions with observers over varying time periods. Examples: The familiar howling of jackals at night, the satisfaction of hunting a notoriously dangerous species, the revulsion for scavengers, etc.

allowing us to incorporate these species in the statistical analysis.

3.3. Charisma score

For this analysis, we assigned values of species' charisma using the functional definition provided by Lorimer (2006, 2007) and his description of the four components of charisma (DETEC, ECON, AESTH and INTELL) that we adapted to fit the group of people considered in this study, namely early European literate settlers and travelers, hereafter referred to as “observers”, acknowledging the fact that the perception of charisma is dependent on the group of people considered (Table 1).

For obvious reasons, we were not able to interview observers regarding their own perception of species charisma. Instead, we used quotes from the written historical literature to capture the relationship of these observers with the large mammal fauna of South Africa. Most citations were extracted from Boshoff and Kerley (2013) and Skead (2007, 2011) and references therein, which provide a detailed summary of historical written records of these large terrestrial mammals, with associated quotes from the historical literature. Additional information were collected opportunistically from other historical sources (e.g. Levaillant, 1790, 1796; Burchell, 1822; Mentzel, 1921). With this method, it is thus the charisma perceptions of tens of historical travelers, naturalists, missionaries and literate pioneers in the historical period that have been collated, using their written descriptions of species as supporting evidence.

For each species, the two authors attributed a score of 0 to 1 to each of the four components of charisma defined in Table 1, on a five point scale (0, 0.25, 0.5, 0.75, 1). Both authors were blind to each other's scores and to the values of B_i identified from the historical dataset. Upon comparisons of the rating, and in case of disagreement, additional evidences from the historical literature were collected and discussed until a consensus was reached. The long experience of the second author (GK) in reading and extracting large mammal occurrence data from historical written material was also fundamental for this exercise. Finally, we averaged the values of the four components to obtain a charisma score (C_i) between 0 and 1, 1 being the highest score, with the underlying assumption that the four components contribute equally to species' charisma. See Table A.2 of Appendix for a detailed justification of the rating of charisma for each species.

3.4. Statistical analysis

The plot of B_i against C_i and of B_i against each of the four components of charisma suggested exponential relationships, so we log-transformed B_i and built a linear regression model (LRM) between $\log(B_i)$ and C_i to test if the taxonomic bias observed in our dataset is positively related to species' charisma. We quantified how much of the variance in the reporting bias index is explained by the charisma index C_i alone using the coefficient of determination R^2 . We also built a model excluding the 7 species that were not reported in the

historical literature despite being historically present in the CFR, to investigate the leverage that these species have on the LRM's R^2 .

We were also interested in knowing the contribution of the four components of C_i - DETEC, ECON, AESTH and INTELL - in explaining the taxonomic bias. We first tested the collinearity among the four covariates (DETEC, ECON, AESTH and INTELL) using 1) Pearson's correlation coefficient and 2) the Variation Inflation Factors (VIFs) (Zuur et al., 2010). The Pearson's correlation coefficient is superior to 0.7 for 4 of the 6 combinations of components possible and the VIFs for two of the four predictors are superior to 3, the threshold suggested in Zuur et al. (2010) above which a covariate should be dropped from the analysis. In our case, with a low sampling size ($n = 37$), even a low collinearity might be problematic and for this reason we decided against using multivariate regression. Instead, we used two different approaches: first, we built four separate LRMs to test the relationship between $\log(B_i)$ and each of the components of charisma, allowing us to assess the performance of each of these models separately and compare them to the LRM built with the charisma index C_i . Then, we used Partial Least Square (PLS) regression to investigate the correlation between $\log(B_i)$ and the four components of charisma together. The PLS regression computes latent variables obtained as the linear combination of predictor variables (DETEC, ECON, AESTH and INTELL) that maximize the explained variance in the dependent variable ($\log(B_i)$). This statistical tool is particularly useful for predicting a dependent variable from a set of collinear predictor variables, especially in cases of small sample size (Carrascal et al., 2009). To evaluate the contribution of each component of charisma in the PLS regression, we calculated the explained variance of each variable for the first PLS-component. We also plotted the PLS circle of correlation that summarizes 1) the correlations (as angles) between predictors, response, and PLS components, and 2) how well the variables are explained by the two PLS components, jointly by the distance of the corresponding point from the origin (or, in other words, its proximity to the correlation 1 circle) and individually by the distance for the projections of this point onto the horizontal and vertical axes. We used the *lme4* (Bates et al., 2014) and *plsdepot* (Sanchez, 2012) packages, respectively, for the LRMs and the PLS regression, in R.3.3.3 (R Core Team, 2017).

4. Results

Eighteen species are over-reported in the historical dataset compared to what would be expected given their historical abundance, 12 species are under-reported and 7 species are not reported at all (Fig. 1, Table 2). African buffalo *Syncerus caffer*, the lion *Panthera leo* and the African elephant *Loxodonta africana* are the most over-represented species in the dataset, being reported ca. 200, 100 and 90 times, respectively, more than expected from their historical frequency. In contrast, the dataset is strongly biased against 15 species (aardvark, aardwolf, African wild cat, blue duiker, bushpig, Cape fox, Cape grysbok, Cape porcupine, common duiker, honey badger, klipspringer, mountain reedbuck, small spotted cat, steenbok and vervet monkey),

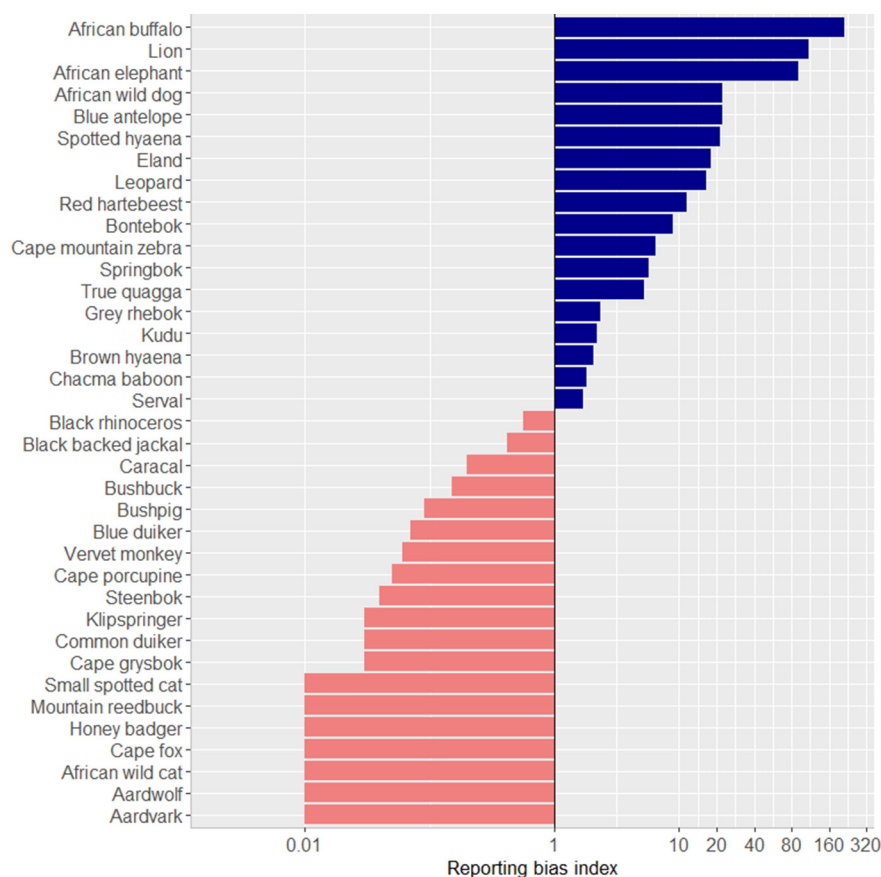


Fig. 1. Taxonomic bias in the historical dataset. Species are ordered by decreasing values of reporting bias index (B_i) and the scale is in log. Dark blue bars correspond to $B_i > 1$, i.e. the species is over-reported compared to what would be expected given its historical abundance (vice versa for light pink bars). Seven species were not represented in the historical dataset despite being historically present. These were given a value of B_i of 0.01, a very low but non-null value, to allow their inclusion in statistical analyses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that are represented less than ten times less than expected from their historical frequency ($B_i < 0.1$). B_i for the most represented species in the dataset (the African buffalo) is four orders of magnitudes higher than the least represented species (Table 2).

The rating of species' charisma by the two authors yielded similar results (in 77% of cases the authors had a disagreement equal to or below 0.25 in the rating of the four components of charisma) and a consensus was successfully reached for all species upon discussion and further consideration of evidences from the historical literature. The resulting charisma index ranges from 0.06 for the less charismatic species (aardwolf *Proteles cristata*, bushpig *Potamochoerus larvatus* and blue duiker *Philantomba monticola*) to 1 for the most charismatic species (African elephant and lion) (Table 2). The African wild dog *Lycan pictus* and the eland *Tragelaphus oryx* are the next most charismatic species, with $C_i > 0.9$.

The LRM shows a strong positive relationship between $\log(B_i)$ and C_i (Fig. 2). The model's R-squared is 0.75, i.e. 75% of the variance in B_i is explained by C_i alone. The power relationship between the two parameters can be summarized in the following way: $B_i = 0.01 + C_i \times e^{8.5}$. The model relating B_i to C_i has a lower AIC and higher R^2 compared to models relating B_i to any of the four components of charisma taken separately (Table 3), suggesting that the model using C_i as a dependent variable is of better quality and provides a better fit to the data. A LRM excluding the seven species that are unreported in the historical dataset returns a R-squared of 0.68, showing that the strength of the relationship between B_i and C_i remains valid when these species are not considered.

The R^2 for the first component of the PLS model is 0.77 (Table 3), i.e. 77% of the total variation is explained by the first latent variable of the PLS model. The length of variable vectors (Fig. 3) and the associated R^2 (Table 3) indicate that INTELL and ECON are the most important variables in predicting B_i , while DETECT is the least important one. The angles between the variable vectors on the PLS circle of correlation are

small, showing that the predictors are highly correlated.

The black rhinoceros *Diceros bicornis*, grey rhebok *Pelea capreolus* and serval *Leptailurus serval* are identified as outliers in the LRM (plot of residuals vs leverage, data not shown). The black rhinoceros is surprisingly sparsely reported given its high charisma index, whereas the two other species are over-reported despite low values of charisma.

5. Discussion

This study provides evidence that a strong taxonomic bias, with up to several order of magnitudes of difference in reporting between some species, exists in biodiversity datasets extracted from historical written sources. We show that species' charisma alone explains 75% of the observed variance, with the most charismatic species being largely over-reported relative to their historical relative frequency. This is, to our knowledge, the first explicit evidence of a positive relationship between taxonomic bias and charisma in a historical biodiversity dataset.

5.1. Limitations

The relevance of summarizing a subjective and complex notion such as charisma with a single quantitative value is debatable, as is our capacity to infer human perception of nature for a group of people operating in a very different sociological context. We attempted to make the analysis as objective and reproducible as possible, by basing our rating on a functional definition of charisma (Lorimer, 2007), informed by views expressed in the historical literature. Al-Abdulrazzak et al. (2012) found that people's perceptions of anecdotal evidences are generally reliable to acquire quantitative data, supporting the idea that using the historical literature to infer observer's perception of nature is relevant for this exercise. It could have been beneficial for the robustness of the approach to have a larger number of people contribute to the

Table 2

Calculation of reporting bias (B_i) and charisma index (C_i) (in bold) for the 37 species of large terrestrial mammals considered in this study. The observed abundance was calculated as the number of occurrences in the historical dataset, multiplied by the mean group size for each species (see text for methods). Historical abundances are estimates from Kerley et al. (2003b). Observed/historical frequencies are obtained by dividing the observed/historical abundance of each species by the total abundance of all species. B_i is the ratio between observed frequency and historical frequency. Note that due to rounding values to the second decimal in the table, the values of B_i are slightly different from what they would be if calculated directly from this table. C_i is the mean of the four components of charisma: Detectability (DETEC), Economic bias (ECON), Aesthetic (AESTH) and Intellectual satisfaction (INTELL). See Table 1 for a description of each component and Table A.2 for rationale of attributed values for each species. *Assigned minimal value due to no observations for this species.

Species		Reporting bias					Charisma				
Vernacular name	Scientific name	Observed abundance	Observed frequency	Historical abundance	Historical frequency	Reporting bias index (B_i)	DETEC	ECON	AESTH	INTELL	Charisma index (C_i)
Aardvark	<i>Orycteropus afer</i>	0	0	2895	0.1	0.01*	0.25	0.25	0.25	0	0.19
Aardwolf	<i>Proteles cristata</i>	0	0	16,886	0.57	0.01*	0.25	0	0	0	0.06
African buffalo	<i>Syncerus caffer</i>	3100	26.82	3746	0.13	213.85	0.75	1	0.75	1	0.88
African elephant	<i>Loxodonta africana</i>	1020	8.83	2906	0.1	90.71	1	1	1	1	1.00
African wild cat	<i>Felis silvestris lybica</i>	0	0	85,462	2.86	0.01*	0	0.25	0.25	0	0.13
African wild dog	<i>Lycan pictus</i>	228	1.97	2622	0.09	22.48	1	1	0.75	1	0.94
Black backed jackal	<i>Canis mesomelas</i>	17	0.15	10,703	0.36	0.42	1	0.5	0	1	0.63
Black rhinoceros	<i>Diceros bicornis</i>	30	0.26	14,041	0.47	0.56	0.5	1	0.75	1	0.81
Blue antelope	<i>Hippotragus leucophaeus</i>	32	0.28	373	0.01	22.18	1	0.5	0.75	0.5	0.69
Blue duiker	<i>Philantomba monticola</i>	18	0.16	83,966	2.81	0.07	0	0	0.25	0	0.06
Bontebok	<i>Damaliscus p. pygargus</i>	232	2.01	6677	0.22	8.99	1	0.75	0.5	0.5	0.69
Brown hyaena	<i>Parahyaena brunnea</i>	27	0.23	3396	0.11	2.06	0.25	0.25	0.75	0.5	0.44
Bushbuck	<i>Tragelaphus scriptus</i>	15	0.13	27,654	0.93	0.15	0.25	0.5	0.25	0	0.25
Bushpig	<i>Potamochoerus larvatus</i>	10	0.09	32,504	1.09	0.09	0	0.25	0	0	0.06
Cape fox	<i>Vulpes chama</i>	0	0	27,942	0.94	0.01*	0	0.25	0.25	0	0.13
Cape grysbok	<i>Raphicerus melanotis</i>	26	0.22	322,977	10.81	0.03	0.25	0	0.25	0	0.13
Cape mountain zebra	<i>Equus z. zebra</i>	180	1.56	7249	0.24	6.43	0.5	0.25	0.75	0.5	0.50
Cape porcupine	<i>Hystrix africae australis</i>	36	0.31	245,175	8.21	0.05	0.25	0.25	0.5	0.5	0.38
Caracal	<i>Caracal caracal</i>	3	0.03	4054	0.14	0.2	0	0.25	0.25	0	0.13
Chacma baboon	<i>Papio ursinus</i>	1920	16.61	271,971	9.11	1.83	1	0.25	0.75	0.5	0.63
Common duiker	<i>Sylvicapra grimmia</i>	25	0.22	283,019	9.48	0.03	0.25	0.25	0.25	0.25	0.25
Eland	<i>Tragelaphus oryx</i>	189	1.64	2736	0.09	17.86	1	1	0.75	1	0.94
Grey rhebok	<i>Pelea capreolus</i>	312	2.7	34,507	1.16	2.35	0.25	0	0.25	0.25	0.19
Honey badger	<i>Mellivora capensis</i>	0	0	2186	0.07	0.01*	0.25	0	0	0.25	0.13
Klipspringer	<i>Oreotragus oreotragus</i>	22	0.19	250,749	8.4	0.03	0.25	0.25	0.25	0.25	0.25
Kudu	<i>Tragelaphus strepsiceros</i>	77	0.67	9186	0.31	2.18	0.75	0.75	0.75	0.75	0.75
Leopard	<i>Panthera pardus</i>	55	0.48	866	0.03	16.42	0.5	1	1	1	0.88
Lion	<i>Panthera leo</i>	750	6.49	1759	0.06	110.19	1	1	1	1	1.00
Mountain reedbuck	<i>Redunca fulvorufula</i>	0	0	2877	0.1	0.01*	0.25	0.25	0.25	0	0.19
Red hartebeest	<i>Alcelaphus buselaphus</i>	1260	10.9	28,339	0.95	11.5	1	0.75	0.5	0.25	0.63
Serval	<i>Leptailurus serval</i>	26	0.22	3938	0.13	1.72	0	0.25	0.5	0	0.19
Small spotted cat	<i>Felis nigripes</i>	0	0	6343	0.21	0.01*	0	0.25	0.25	0	0.13
Spotted hyaena	<i>Crocuta crocuta</i>	220	1.9	2666	0.09	21.33	1	0.75	0.75	0.75	0.81
Springbok	<i>Antidorcas marsupialis</i>	1400	12.11	63,303	2.12	5.72	1	0.5	0.5	0.75	0.69
Steenbok	<i>Raphicerus campestris</i>	34	0.29	307,173	10.29	0.04	0.5	0.25	0.25	0	0.25
True quagga	<i>Equus q. quagga</i>	144	1.25	7087	0.24	5.26	1	1	0.75	0.75	0.88
Vervet monkey	<i>Chlorocebus pygerythrus</i>	150	1.3	806,645	27.01	0.06	1	0	0.25	0	0.31

interpretation of these anecdotes. There are however only a handful of people who have a sufficient comprehension and experience of the South African historical zoological literature to perform this exercise. Overall, the method to rate species' charisma only involved little discrepancy between the two authors, and these were dealt with a transparent approach, constantly referring to the historical literature, and providing all details of the charisma scoring in Table A.2 of the appendices. We hope that this demonstrates our effort in making the study

as robust, transparent and reproducible as possible. While the values of charisma presented in this study are only applicable to a particular group of people and reflect their comments on the various species, we hope that the method itself will be tested in other spatio-temporal contexts to test its effectiveness.

While previous studies investigating taxonomic bias in biodiversity data considered higher taxonomic groups (e.g. Clark and May, 2002; Troudet et al., 2017), this analysis is restricted to large terrestrial

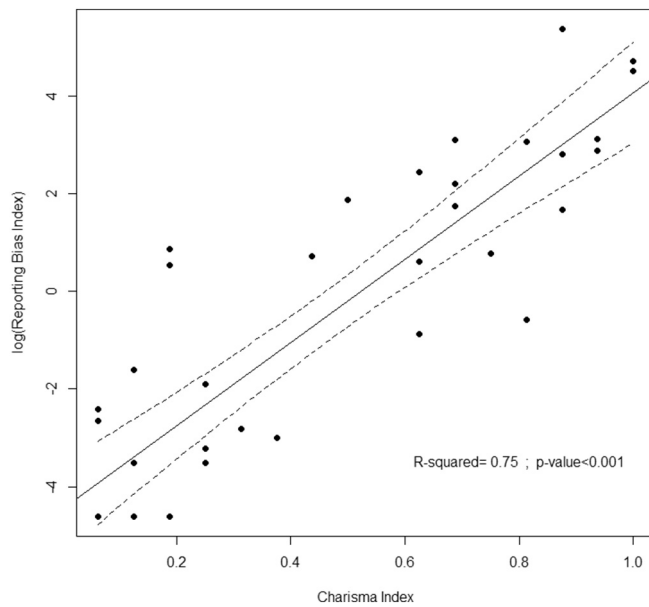


Fig. 2. Reporting bias index (B_i) plotted against the charisma index (C_i) for the 37 species considered in the analysis. The regression line (plain line) and 95% confidence intervals (dotted lines) are the result of a linear regression model between C_i and $\log(B_i)$ (formula: $\log(B_i) = -4.5 + 8.5 \times C_i$). B_i is log-transformed, indicating a power relationship between B_i and C_i .

Table 3

Summary results from the Linear Regression Models (LRM) and Partial Least-Square (PLS) regression. Five LRMs were built to test the relationships between the reporting bias index B_i and the four components of C_i separately (detectability DETEC, economic bias ECON, aesthetic AESTH and intellectual satisfaction INTELL) and the relationship between B_i and the charisma index C_i , defined as the mean of these four components. The PLS regression included the four components of charisma as dependent variables to explain B_i . Only the results for the first component of the PLS regression are presented, as it explains most of the total variance, and to allow the comparison with the LRM that uses C_i as the dependent variable. ***p-value < 0.001.

LRM				PLS regression	
Variable	R ²	p-value	AIC	Variable	R ² (first component)
DETEC	0.51	***	168	DETEC	0.67
ECON	0.57	***	163	ECON	0.83
AESTH	0.68	***	153	AESTH	0.77
INTELL	0.60	***	160	INTELL	0.86
C_i	0.75	***	142	Global	0.77

mammals, a highly charismatic taxonomic group that is substantially overrepresented in conservation research and biodiversity datasets (Clark and May, 2002; Troudet et al., 2017). This must be taken into account before drawing more general conclusions from these results. Other animal taxa that are under-represented in biodiversity data (e.g. reptiles, amphibians, and invertebrates in general) are also seldom cited in the historical written literature (Rookmaaker, 1989), which makes it difficult to carry out similar studies for these organisms.

Despite a strong relationship between the reporting bias and charisma indexes, outliers in the analysis were identified (black rhinoceros, serval and grey rhebok). The results for these species suggest that factors other than charisma, possibly species-specific and hence idiosyncratic, are important in predicting taxonomic bias. Alternatively, these outliers may be the consequence of limitations of the analysis (e.g. some aspects of the observers' perception of these species was not well captured in the sample of historical accounts considered in this study; these species' habitat was under or over-sampled by observers, which introduced a bias in the calculation of B_i), but these different factors are

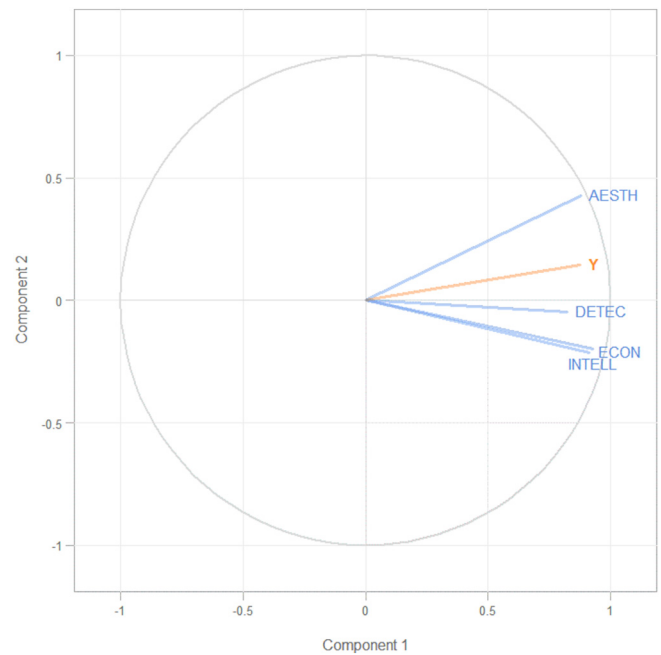


Fig. 3. PLS circle of correlations plot illustrating relationships between predictor variables (blue), reporting bias index B_i (orange) and PLS components. The angles between the lines indicate the correlation between the different variables (the smaller the angle, the higher the correlation). The closer the variable vector is to the circle of correlation 1, the more important this variable is in predicting B_i . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

difficult to assess. However, the value of the regression between B_i and C_i remains strong despite these outliers, suggesting that, overall, the hypothesis is well supported by the data.

5.2. Implications for the analysis of long-term biodiversity data

This study shows that taxonomic bias does not only affect the level of reporting between higher taxonomic groups as previously demonstrated (Boakes et al., 2016; Troudet et al., 2017), but may also exist within an apparently homogeneous group such as large terrestrial South African mammals. In other words, studies that consider closely-related species should not be spared from investigating the existence and implications of taxonomic bias in their analyses.

In our dataset, seven species (i.e. 18.9% of the dataset) were not reported in historical accounts despite being historically present (Kerley et al., 2003b). This is not unexpected given their behaviour and habitat: six are small, non-threatening nocturnal carnivores/insectivores, while the single ungulate is a rugged habitat specialist with a marginal occurrence in the CFR. These species also have low charisma index ($C_i < 0.2$), and their inclusion in the statistical analysis strengthened the relationship found between the reporting bias and charisma. Surprisingly, given its rarity and early extinction, the blue antelope achieved the 5th highest reporting bias (Fig. 1). This positive reporting bias might be explained by the value that people put on rarity (Angulo and Courchamp, 2009), supporting modern studies on the valuing of African wildlife trophies (Johnson et al., 2010). Kerley et al. (2009) point out that the blue antelope was valued as a 'curiosity' by collectors by virtue of its rarity. These species illustrate the pitfalls of the biases in historical record reporting, as lack of (or over) reporting does not necessarily indicate absence (or abundance) of species. We suggest that the reporting bias index could be used as a correction for the relative abundance of species in analyses of long-term changes in species composition based on these data.

In general, the existence of a strong taxonomic bias in historical

occurrence data risks compromising conclusions on changes in past communities and composition of earlier ecosystems. Unfortunately, the taxonomic bias in long-term biodiversity datasets cannot be corrected with additional contemporary data collection but identifying it can improve the interpretation of analyses based on these data. A direct measure of taxonomic reporting bias requires estimates of species' historical relative frequency. If this information is not available, identifying correlates may instead prove useful to predict which taxa are more or less likely to be reported. In contrast with previous studies (Royle and Nichols, 2003; Boakes et al., 2016), we do not identify abundance and detectability as the best predictors of taxonomic bias in historical occurrence records. The reporting bias index is not equal to one for all species (which would be expected if abundance explained the difference in reporting) and of the four components considered in the study, detectability explains the least variance in the data. Other parameters such as the economic costs and benefits and the emotions experienced by observers in their interactions with species seem to be important predictors of the level of reporting of species. Overall, a multivariate measure such as species' charisma that combines these different parameters better fits the bias in reporting, and may be a useful predictor of taxonomic bias in historical biodiversity datasets in general.

5.3. Charisma, a timeless concept?

Through their descriptions of the environment and the animals they encountered, historical observers provide a glimpse of the attitude of humans towards nature in the early colonial history in South Africa. Overall, the concept of nonhuman charisma, even if only recently defined in the conservation literature, seems to have been stable through time. Big African savanna mammals, and the big cats in particular, are particularly appealing flagship species for conservation, being highly favoured by the public. A study assessing the relative charisma of different mammals based on an online survey on 1500 people from five continents found that six of the ten top ranking species were members of the Felidae family, and that the African elephant was ranked second in three of the five continents sampled (Macdonald et al., 2015). It is also well documented that tourists only appreciate a small proportion of mammal biodiversity (Kerley et al., 2003b), and a study on tourists' preferences during game drives in a South African private protected area found that the most appreciated species were the lion, leopard and elephant (Maciejewski and Kerley, 2014), three species that are rated high in historical charisma ($C_i > 0.88$). It is interesting to note that species that form the “Big Five” - the lion, leopard, elephant, buffalo and rhinoceros, originally considered to be the most dangerous species to hunt in Africa, but now the major wildlife drawcard for the tourism market (Kerley et al., 2003a) - all have a high historical charisma, highlighting the consistency in preferences for these species in historical and modern times. However, we also find discrepancies, e.g. the buffalo's and the eland's historical charisma is very high while these species are both low in the preference of modern tourists (Maciejewski and Kerley, 2014). Modern tourists use such large ungulates for recreational purposes (game-viewing), while early European observers had a more utilitarian use, hunting species for the resources they provide and as trophies, which may explain higher charisma values for these species in early history.

As human use of large mammal wildlife shifts from a utilitarian to a recreational perspective, there seems to remain similarities in the feelings experienced during the chase and hunt of big game and the observation of wild animals in protected areas, as suggested by this quote from St J. MacDonal, visiting the Kruger National Park in 1954: “The watchful experience I have from a car compares with any thrill I experience awaiting a driven tiger, a rogue elephant, or closing in with a tracked bison or tsine [Malayan wild ox, *Bos banteng*]” (MacDonald, 1954).

6. Conclusion

Our analysis confirms that the correlation between societal preference (here approximated by species' perceived charisma) and taxonomic bias is as relevant in early historical as in modern times (Troudet et al., 2017). Though this study does not allow us to clarify the causality issue, it suggests that species' charisma is a good predictor of taxonomic bias in historical biodiversity datasets and that using this concept to predict which taxa are over- or under-reported could allow for better interpretation of analyses based on long-term biodiversity datasets. We show that taxonomic bias is not only a modern issue. As it influences our understanding of global biodiversity patterns over time, and the development of appropriate conservation plans, it is crucial that conservation scientists acknowledge this bias and find ways to address it in current and future biodiversity studies.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.biocon.2018.04.036>.

Funding

This work was supported by a Claude Leon Postdoctoral Fellowship 2017-2018.

References

- Al-Abdulrazzak, D., Naidoo, R., Palomares, M.L.D., Pauly, D., 2012. Gaining perspective on what we've lost: the reliability of encoded anecdotes in historical ecology. *PLoS One* 7, e43386.
- Angulo, E., Courchamp, F., 2009. Rare species are valued big time. *PLoS One* 4, e5215.
- Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version, 1. pp. 1–23.
- Boakes, E.H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D.B., Haklay, M., 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Sci. Rep.* 6, 1–11.
- Boshoff, A.F., Kerley, G.I.H., 2010. Historical mammal distribution data: how reliable are written records? *S. Afr. J. Sci.* 106, 26–33.
- Boshoff, A.F., 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 University.
- Boshoff, A., Landman, M., Kerley, G.I.H., 2016. Filling the gaps on the maps: historical distribution patterns of some larger mammals in part of southern Africa. *Trans. R. Soc. S. Afr.* 71, 1–65.
- Burchell, W.J., 1822. *Travels in the Interior of Southern Africa*. Longman, Hurst, Rees, Orme, and Brown.
- Carrascal, L.M., Galván, I., Gordo, O., 2009. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* 118, 681–690.
- Clark, J.A., May, R.M., 2002. Taxonomic bias in conservation research. *Science* 297, 191.
- Colvin, I., 1912. *The Cape of Adventure*. TC and EC Jack, Edinburgh.
- Ducarme, F., Luque, G.M., Courchamp, F., 2013. What are “charismatic species” for conservation biologists. *BioSciences Master Rev.* 10, 1–8.
- Feeley, K.J., Stroud, J.T., Perez, T.M., 2017. Most ‘global’ reviews of species' responses to climate change are not truly global. *Divers. Distrib.* 23, 231–234.
- Feldhamer, G., Whittaker, J., Monty, A.-M., Weickert, C., 2002. Charismatic mammalian megafauna: public empathy and marketing strategy. *J. Pop. Cult.* 36, 160–167.
- Goodwin, H.J., Leader-Williams, N., 2000. Tourism and protected areas-distorting conservation priorities towards charismatic megafauna? In: *Conservation Biology Series* - Cambridge, pp. 257–276.
- Heywood, V.H. (Ed.), 1995. *Global Biodiversity Assessment*. University of Reading.
- Johnson, P.J., Kinsky, R., Loveridge, A.J., Macdonald, D.W., 2010. Size, rarity and charisma: valuing African wildlife trophies. *PLoS One* 5, e12866.
- Kerley, G.I.H., Geach, B.G., Vial, C., 2003a. Jumbos or bust: do tourists' perceptions lead to an under-appreciation of biodiversity? *S. Afr. J. Wildl. Res.* 33, 13–21.
- Kerley, G.I.H., Pressey, R.L., Cowling, R.M., Boshoff, A.F., Sims-Castley, R., 2003b. Options for the conservation of large and medium-sized mammals in the Cape Floristic Region hotspot, South Africa. *Biol. Conserv.* 112, 169–190.
- Kerley, G.I.H., Sims-Castley, R., Boshoff, A.F., Cowling, R.M., 2009. Extinction of the blue antelope *Hippotragus leucophaeus*: modeling predicts non-viable global population size as the primary driver. *Biodivers. Conserv.* 18, 3235–3242.
- Leader-Williams, N., Dublin, H.T., 2000. Charismatic megafauna as ‘flagship species’. In: *Conservation Biology Series* - Cambridge, pp. 53–84.
- Levaillant, F., 1790. *Travels Into the Interior Parts of Africa: By the Way of the Cape of Good Hope in the Years 1780, 81, 82, 83, 84, and 85*. G. G. J. and J. Robinson.
- Levaillant, 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 Printed for.
- Lorimer, J., 2006. Nonhuman charisma: which species trigger our emotions and why? *ECOS* 27, 20–27.
- Lorimer, J., 2007. Nonhuman charisma. *Environ. Plan. D Soc. Space* 25, 911–932.

- MacDonald, S.J., 1954. Wildlife preservation, the Kruger National Park, an example. *J. Bombay Nat. Hist. Soc.* 52, 1–4.
- Macdonald, E.A., Burnham, D., Hinks, A.E., Dickman, A.J., Malhi, Y., Macdonald, D.W., 2015. Conservation inequality and the charismatic cat: *Felis felis*. *Glob. Ecol. Conserv.* 3, 851–866.
- Maciejewski, K., Kerley, G.I.H., 2014. Understanding tourists' preference for mammal species in private protected areas: is there a case for extralimital species for ecotourism? *PLoS One* 9, e88192.
- McKinney, M.L., 1999. High rates of extinction and threat in poorly studied taxa. *Conserv. Biol.* 13, 1273–1281.
- Mentzel, O.F., 1921. A geographical and topographical description of the Cape of Good Hope - with translation of original t.p.: a complete and authentic geographical and topographical description of the famous and (all things considered) remarkable African Cape of Good Hope / by O.F. Mentzel. Glogau: C.F. Günther, 1785–87. Van Riebeeck Society, Cape Town.
- Mihoub, J.-B., Henle, K., Titeux, N., Brotons, L., Brummitt, N.A., Schmeller, D.S., 2017. Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures. *Sci. Rep.* 7, 41591.
- Myers, N., 1983. A priority-ranking strategy for threatened species? *Environmentalist* 3, 97–120.
- Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A., Kent, J., 2000. Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858.
- 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.
- R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rangarajan, M., 2003. Parks, politics and history: conservation dilemmas in Africa. *Conserv. Soc.* 1, 77.
- Rookmaaker, L.C., 1989. *The Zoological Exploration of Southern Africa 1650–1790*. Taylor & 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. S. Afr.* 62, 55–198.
- Royle, J.A., Nichols, J.D., 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* 84, 777–790.
- Sanchez, G., 2012. Package 'plsdepot'. *Partial Least Squares (PLS) Data Analysis Methods*, v. 0.1.1. pp. 17.
- Skead, C.J., 1980. *Historical Mammal Incidence in the Cape Province, Vol. 1: The Western and Northern Cape*. Department of Nature and Environmental Conservation of the Provincial Administration of the Cape of Good Hope, Cape Town.
- Skead, C.J., 2007. In: Boshoff, A., Kerley, G.I.H., Lloyd, P. (Eds.), *Historical Incidence of the Larger Land Mammals in the Broader Eastern Cape*. Centre for African Conservation Ecology, Nelson Mandela Metropolitan University.
- Skead, C.J., 2011. In: Boshoff, A., Kerley, G.I.H., Lloyd, P. (Eds.), *Historical Incidence of the Larger Land Mammals in the Broader Western and Northern Cape*. Centre for Conservation Ecology, Nelson Mandela Metropolitan University.
- 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. *Biodivers. Conserv.* 9, 1441–1466.
- Szabo, J.K., Vesk, P.A., Baxter, P.W., Possingham, H.P., 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. *Ecol. Appl.* 20, 2157–2169.
- Tingley, M.W., Beissinger, S.R., 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends Ecol. Evol.* 24, 625–633.
- Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* 7 (1), 14.
- Turvey, S.T., Crees, J.J., Di Fonzo, M.M.I., 2015. Historical data as a baseline for conservation: reconstructing long-term faunal extinction dynamics in Late Imperial-modern China. *Proc. R. Soc. B Biol. Sci.* 282, 20151299.
- Van Sittert, L., 2005. Bringing in the wild: the commodification of wild animals in the Cape Colony/Province c. 1850–1950. *J. Afr. Hist.* 46, 269–291.
- Verissimo, D., MacMillan, D.C., Smith, R.J., 2011. Toward a systematic approach for identifying conservation flagships. *Conserv. Lett.* 4, 1–8.
- Western, D., 1987. Africa's elephants and rhinos: flagships in crisis. *Trends Ecol. Evol.* 2, 343–346.
- Willis, K.J., Araujo, M.B., Bennett, K.D., Figueroa-Rangel, B., Froyd, C.A., Myers, N., 2007. How can a knowledge of the past help to conserve the future? Biodiversity conservation and the relevance of long-term ecological studies. *Philos. Trans. R. Soc. B* 362, 175–187.
- Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems: data exploration. *Methods Ecol. Evol.* 1, 3–14.