Made available by Hasselt University Library in https://documentserver.uhasselt.be

Reviewing taxonomic bias in a megadiverse country: primary biodiversity data, cultural salience, and scientific interest of South African animals Non Peer-reviewed author version

PHAKA, Fortunate; VANHOVE, Maarten; du Preez, Louis H. & HUGE, Jean (2022) Reviewing taxonomic bias in a megadiverse country: primary biodiversity data, cultural salience, and scientific interest of South African animals. In: ENVIRONMENTAL REVIEWS, 30 (1), p. 39 -49.

DOI: 10.1139/er-2020-0092 Handle: http://hdl.handle.net/1942/37210

1	Reviewing taxonomic bias in a megadiverse country: primary biodiversity data, cultural
2	salience, and scientific interest of South African animals
3	
4	Fortunate M. Phaka ^{1,2*} , Maarten P. M. Vanhove ^{2, 3, 4, 5} , Louis H. du Preez ^{1,6} , Jean Hugé ^{7, 8, 9, 10}
5	
6	¹ African Amphibian Conservation Research Group, Unit for Environmental Sciences and
7	Management, North-West University, Private Bag X6001, Potchefstroom, North-West, South
8	Africa, 2520
9	² Hasselt University, Centre for Environmental Sciences, Research Group Zoology:
10	Biodiversity and Toxicology, Diepenbeek, Limburg, Belgium
11	³ Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, South
12	Moravia, Czech Republic
13	⁴ Zoology Unit, Finnish Museum of Natural History, University of Helsinki, Helsinki, Uusimaa,
14	Finland
15	⁵ Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, University
16	of Leuven, Leuven, Flemish Brabant, Belgium
17	⁶ South African Institute for Aquatic Biodiversity, Somerset Street, Makhanda, Eastern Cape,
18	South Africa
19	⁷ Hasselt University, Centre for Environmental Sciences, Research Group Environmental
20	Biology, Diepenbeek, Limburg, Belgium
21	⁸ Open University of the Netherlands, Heerlen, Limburg, the Netherlands
22	⁹ Systems Ecology & Resource Management Unit, Université Libre de Bruxelles, Brussels,
23	Belgium
24	¹⁰ Biology Department, Vrije Universiteit Brussel, Brussels, Belgium

- 25 * Corresponding author: email: mafetap@gmail.com | Phone: +27 71 436 6065 / +32 489 44
- 26 02 83

Word count: 8407 (excluding Abstract)

27 Abstract

28 Taxonomic bias, resulting in some taxa receiving more attention than others, has been 29 shown to persist throughout history. Such bias in primary biodiversity data needs to be 30 addressed as the data is vital to environmental management. This study reviews taxonomic 31 bias in South African primary biodiversity data obtained from the Global Biodiversity 32 Information Facility (GBIF). The focus is specifically on animal classes, and regression 33 analysis is used to assess the influence of scientific interest and cultural salience on 34 taxonomic bias. A higher resolution analysis of the two explanatory variables' influence on 35 taxonomic bias is conducted using a Generalised Linear Model on a subset of herpetofaunal 36 families from the focal classes. Furthermore, the potential effects of cultural salience and 37 scientific interest on a taxon's extinction risk are investigated. The findings show that 38 taxonomic bias in South Africa's primary biodiversity data has similarities with global scale 39 taxonomic bias. Among animal classes, there is strong bias towards birds while classes such 40 as Polychaeta and Maxillopoda are underrepresented. Cultural salience has a stronger 41 influence on taxonomic bias than scientific interest. It is, however, unclear how these explanatory variables may influence the extinction risk of taxa. We recommend that 42 43 taxonomic bias can be reduced if primary biodiversity data collection has a range of targets 44 that guide (but do not limit) accumulation of species occurrence records per habitat. Within 45 this range, a lower target of species occurrence records accommodates species that are 46 difficult to detect. The upper target means occurrence records for any species are less 47 urgent but nonetheless useful and thus data collection efforts can focus on species with 48 fewer occurrence records.

49

- 50 Keywords; Biodiversity hotspot, Conservation, Data collection, Surrogate species,
- 51 Taxonomic chauvinism

52	Background
53	Primary biodiversity data reflects knowledge and practices in the study of biodiversity
54	(Troudet et al. 2017) and consists of species occurrence records (Soberón and Peterson
55	2004). These occurrence records include details about taxonomic information,
56	collection/observation date and location (Ball-Damerow et al. 2019). Accumulating primary
57	biodiversity data is an important step towards biodiversity conservation policy (Flemons et
58	al. 2007) and it is also vital to biodiversity research (Lira-Noriega 2007). Making primary
59	biodiversity data publicly available is an important requirement for biodiversity research
60	and planning (Huang et al. 2013). An organisation called Global Biodiversity Information
61	Facility (GBIF) facilitates the sharing of and access to accumulated primary biodiversity data
62	(Edwards 2004) and in 2012 GBIF was said to host the largest open access biodiversity
63	database in the world (Boyd and Crawford 2012). GBIF is rapidly growing and at the time of
64	writing this article, the GBIF network consisted of 101 countries and international
65	organisations dedicated to advancing open access primary biodiversity data (GBIF 2020b).
66	
67	The biodiversity data making up the GBIF (2020a) dataset consists of contributions from
68	data publishers which include academic institutions, museums, herbaria, non-governmental
69	organisations, and citizen science projects such as iNaturalist (www.inaturalist.org) and
70	AntWeb (www.antweb.org). Georeferenced data from these data publishers is collected by
71	individuals with varying levels of biodiversity knowledge including scientists and amateur
72	biodiversity enthusiasts (GBIF 2020a). The GBIF dataset is highly cited in scientific articles

- 73 (Ball-Damerow et al. 2019), thus suggesting that researchers consider GBIF to be a reliable
- 74 biodiversity database. There are also many other biodiversity databases through which
- 75 primary biodiversity data is published and accessed: these include Integrated Digitized

76 Biocollections and The Barcode of Life Data System (Ball-Damerow et al. 2019). Biodiversity 77 databases are not exempt from bias. Large disparities in the number of occurrence records 78 for one taxon, in comparison to other taxa in a biodiversity dataset, highlights a taxonomic 79 bias. This taxonomic bias is a global phenomenon that has persisted for many years (Russell 80 1984; Ponder 1992; Troudet et al. 2017; Gordon et al. 2020). Understanding taxonomic bias 81 in primary biodiversity data assists with addressing knowledge gaps and informing policy 82 (Donaldson et al. 2016) as these data are important for biodiversity research and 83 management (Flemons et al. 2007; Lira-Noriega 2007; Huang et al. 2013). Taxonomic bias 84 can lead to underestimation of the extinction risk and future threats for underrepresented taxa (McKinney 1999), may limit understanding of how natural systems are affected by 85 86 anthropogenic disturbances (Feeley et al. 2016), and could also divert conservation 87 resources away from taxa that urgently need them (Seddon et al. 2005). 88

89 Biased representation of taxa in primary biodiversity data may result from intrinsic features, 90 such as abundance, remoteness and behaviour, which can make it difficult to obtain the 91 occurrence records of some species. More extinction resistant (common species) tend to be 92 recorded first (McKinney 1999). Primary biodiversity data is generally biased towards 93 species that are easy to identify (Boakes et al. 2016) or locally abundant (Royle and Nichols 94 2003). Intrinsic features are, however, not solely responsible for taxonomic bias and the 95 cause of this bias is not fully understood (Troudet et al. 2017). It has been hypothesized that 96 taxonomic bias in primary biodiversity data is influenced by either scientific interest (Pawar 97 2003) or societal interest (Wilson et al. 2007). Taxonomic bias exists in the interests of both 98 science (Clark and May 2002; Di Marco et al. 2017) and society (Ducarme et al. 2013). 99 Studies, such as this current one, aim to understand which of the biases, in the interests of

Environmental Reviews

100	science and society, lead to taxonomic bias in primary biodiversity data accumulation.
101	Scientific interest can be deduced from scientific articles (Troudet et al. 2017), as these
102	articles provide an idea of which taxa the scientific community is dedicating its resources to.
103	Societal interest can be quantified from frequency of words in web pages (Wilson et al.
104	2007). The frequency, or number of times, that words (e.g., taxon names) occur in a large
105	body of text such as web pages can be used as a measure for cultural salience (Correia et al.
106	2016; Davies et al. 2019). This cultural salience denotes the cultural visibility and profile of
107	species (Correia et al. 2016; Davies et al. 2019), or the popularity of a species which reflects
108	the interactions of cultural value-practice systems with that species' traits (Ducarme et al.
109	2013). According to Correia et al. (2017), inferring cultural salience from web pages as large
110	bodies of text relies on the assumption that content on the World Wide Web is a reflection
111	of the interests of the citizenry generating it. There is evidence supporting this assumption
112	that internet activity reflects societal interests (Funk and Rusowsky 2014; Schuetz et al.
113	2015; Troumbis 2017; Kim et al. 2018) . When conducting web page searches, scientific and
114	vernacular names of taxa can be used interchangeably as keywords (Jarić et al. 2016). There
115	are high correlations between scientific and vernacular species name search results at both
116	global and country level (e.g. in . Australia, Brazil, Indonesia, Spain, Tanzania and United
117	States of America) regardless of lingual and cultural differences (Correia et al. 2017). High
118	correlations are also found between the Google search engine results for scientific and
119	English names of diurnal birds of prey, carnivores, and primates (Jarić et al. 2016).
120	
121	Investigations of taxonomic bias generally focus on its causes and seldom discuss what
122	would be considered an ideal representation of a taxon. This ideal representation of a taxon

123 would provide guiding and non-limiting targets for primary biodiversity data collection.

124 Proposing target species occurrence records to guide primary biodiversity data collection 125 would also be cognisant of the sample size requirements of biodiversity data applications. 126 When using biodiversity data for regression analysis, at least ten subjects per variable are required for accurate models (Harrell 2001). High accuracy species distribution models can 127 128 be created using samples of five, ten or 25 (Hernandez et al. 2006). For African taxa in 129 particular, accurate species distribution models can be developed with at least 14 130 occurrence records for species with a limited distribution range, and 25 records for widely 131 distributed species (Van Proosdij et al. 2016). In biostatistics, a minimum sample size of 30 is 132 considered sufficient for the design of field studies (Cohen and Cohen 1995). To 133 accommodate these varied sample size requirements of biodiversity data applications 134 would require the target occurrence records per species to be made up of a range of 135 numbers. A range of target sample sizes that would be considered an ideal representation 136 of a taxon in primary biodiversity data is also better suited to the varying levels of difficulty 137 in collecting species occurrence records. The ease with which species occurrence records 138 can be collected is often affected by species' abundance, habits, habitat type and 139 accessibility of habitats.

140

Biodiversity research focus is misaligned with global biodiversity distribution (Griffiths and Dos Santos 2012; Di Marco et al. 2017). Considering that South Africa is a megadiverse country (Mittermeier, 1988; Mittermeier et al. 1997), it is important to understand the extent of biodiversity knowledge and also to investigate factors influencing primary biodiversity data accumulation in order to inform research and planning. This current study seeks to (1) quantify taxonomic bias in the primary biodiversity data of South African animal taxa and make suggestions of the ideal representation of a taxon required in order to lessen

148 bias. Based on hypotheses put forward in previous research (Pawar 2003; Wilson et al

149 2007), we (2) test the likely influence of cultural salience and scientific interest on

150 taxonomic bias in primary biodiversity data. We further (3) assess how cultural salience and

151 scientific interest compare with a taxon's extinction risk.

152 **Review Approach**

153 This assessment of taxonomic bias in South Africa focuses on the timespan from 1998, when 154 the country's National Environmental Management Act (107 of 1998) was promulgated, and 155 the commencement of this study in 2020. The act sought to increase biodiversity monitoring 156 and access. The highest yearly species occurrence records for South African species (ranging 157 between 20,271 and 1,6 million) started being submitted to GBIF in the third year of this act 158 being in effect. The dataset containing species occurrence records of South African animals 159 was downloaded from the GBIF database on 21 February 2020, and the search parameters 160 used to obtain this dataset are viewable on doi.org/10.15468/dl.5upuwl and in 161 Supplementary Material 1. Not all taxon names used in GBIF datasets represent natural 162 groupings, but those names are used here verbatim for the sake of consistency and 163 comparability. Occurrence statistics are computed based on species occurrence records 164 from the GBIF dataset and the number of known animal species distributed within South 165 Africa. The numbers of known animal species are obtained from the South African Animal 166 Checklist maintained by the South African GBIF node (SANBI Biodiversity Advisor 2020). A 167 ratio of GBIF occurrence records (or species occurrence records submitted to GBIF) to 168 number of known species is used to determine the average number of times each species 169 had their occurrence records submitted to GBIF. Medians of GBIF occurrence records are

used for their robustness to outliers. Absolute deviation around the median (i.e., median
absolute deviation) gives a measure of variability (Troudet et al. 2017).

172

173 Since no guidelines exist for the ideal number of species occurrence records required for 174 taxa to be considered sufficiently represented, we deem it necessary to suggest a range of 175 guiding targets. Lower and upper targets of 10 and 30 occurrence records per species per 176 habitat respectively can serve to guide biodiversity data collection. This range of targets is 177 based on sample sizes that are suitable for various biodiversity data applications. The targets also consider the intrinsic traits of the various species, with the lower target being 178 179 suitable for collecting data about species that are difficult to detect, and the upper target 180 being more suitable for abundant species that are easier to detect. Should the upper target 181 be reached for any species, then researchers can consider re-directing data collection 182 resources to species with occurrence records that are lower than the suggested targets in a 183 habitat. With such guiding targets, biodiversity data collection can be orientated towards 184 lessening persisting taxonomic bias while ensuring there is sufficient primary biodiversity 185 data for research and management.

186

187 Cultural salience and scientific interest are investigated as explanatory variables for 188 taxonomic bias in primary biodiversity data. Regression analysis is used to understand the 189 relative influence of the two explanatory variables and their interactions on GBIF occurrence 190 records. We use scatterplots and Pearson's correlation to investigate correlations between 191 the dependent variable (GBIF occurrence records) and the two explanatory variables. 192 Cultural salience, in the South African context is determined by the frequency of focal taxa 193 names in the South African web corpus (or body of text contained in web pages). The

advanced search option on the Google search engine,

195 (https://www.google.com/advanced search?), was used to search for the exact scientific 196 names of each class (e.g. "Amphibia") and the search was narrowed by region to South 197 Africa (Supplementary Material 1). Google search results include a numeric value which 198 represents cultural salience or the approximate number of times a search term appears in 199 the South African web corpus. Based on previous studies (Funk and Rusowsky 2014; Schuetz 200 et al. 2015; Troumbis 2017; Kim et al. 2018) we assume that the South African web corpus is 201 a general reflection of the interests of South Africans that are generating it. We say 'general 202 reflection' as acknowledgement that South Africa is a diverse country with varied access to 203 the internet. Thus, interests of people that voluntarily avoid using the internet and those in 204 remote areas without internet access, will not be reflected in the South African web corpus. 205 A more representative measure of South African public interest would require an extensive 206 survey of the country's citizenry. Scientific interest is quantified by the number of scientific 207 articles published within this study's focal timeframe with their topic being South African 208 animal taxa. This scientific interest data was obtained from Web of Science (WoS) 209 (<u>http://apps.webofknowledge.com/</u>). The WoS search query (Supplementary Material 1) 210 was as follows "Class" OR "Family" AND "South Africa*"; e.g. TS=("Amphibia" OR 211 "Arthroleptidae" OR "Brevicipitidae" OR "Bufonidae" OR "Heleophrynidae" OR 212 "Hemisotidae" OR "Hyperoliidae" OR "Microhylidae" OR "Phrynobatrachidae" OR "Pipidae" 213 OR "Ptychadenidae" OR "Pyxicephalidae" OR "Rhacophoridae") AND TS=("South Africa*"). 214 The result of this search includes a numeric value which represents scientific interest or the 215 number of scientific articles with a South African animal taxon as their topic (specifically 216 from journals indexed by WoS).

217

Page 12 of 49

218 For a higher resolution investigation into the effects of explanatory variables on taxonomic 219 bias, we additionally analysed a subsample consisting of herpetofauna (amphibians and 220 reptiles). We chose herpetofauna for the subsample as they are generally underrepresented 221 in wildlife research and management literature (Christoffel and Lepczyk 2012), are 222 negatively perceived by the general public (Reimer et al. 2013; Tarrant et al. 2016) and their 223 populations are experiencing global declines (McCallum 2007). A Generalized Linear Model 224 (GLM) fitted using a negative binomial distribution is employed to analyse influence of the 225 two explanatory variables on the number of GBIF occurrence records for each family within 226 the subsample. The validity of the models is checked by testing homogeneity of residuals 227 when plotting the values of residuals against predicted values. Outliers were excluded as 228 they negatively impacted the model's resolution. Identification of outliers was achieved 229 using the Interquartile rule; Q1 – 1.5 x Interquartile Range or above Q3 +1.5 × Interquartile 230 Range. Similar to the methods for the main sample mentioned above, cultural salience for 231 the subsample was obtained through a search of exact family names (e.g. "Arthroleptidae") 232 on Google's advanced search page with results narrowed by region to South Africa only 233 (Supplementary Material 1). The subsample's scientific interest was retrieved from WoS 234 using the following query (Supplementary Material 1) "Family" OR "Genus" AND "South 235 Africa*": e.g. TS=("Arthroleptidae" OR "Leptopelis" OR "Arthroleptis") AND TS=("South 236 Africa*"). For increased accuracy of the subsample's regression analysis, families with less 237 than ten GBIF occurrence records were excluded from the analysis as a minimum of ten 238 subjects per variable is necessary for greater accuracy of regression modelling (Harrell 239 2001).

240

241	This study's analysis is extended to the likely influence of the two explanatory variables on
242	conservation status by comparing cultural salience and scientific interest with threat status
243	of South African animals that are on the IUCN Red List of threatened species (IUCN 2020).
244	This IUCN (2020) Red List has the total number species in all threatened categories per
245	country organised by taxonomic groupings (i.e. birds, mammals, molluscs, fishes, reptiles,
246	amphibians). The seventh IUCN (2020) grouping of animal taxa, labelled "other
247	invertebrates", was excluded from analysis due to uncertainty about which taxa are
248	included in the grouping.
249	
250	Twelve classes, from eight phyla, with occurrence records submitted to GBIF during this
251	study's timeframe were not listed on the South African Animal Checklist (SANBI Biodiversity
252	Advisor 2020), making it impossible to compute their ratios for occurrence records to known
253	species (Table 1). Six classes, from five phyla, listed on the South African Animal Checklist
254	(SANBI Biodiversity Advisor 2020) do not have occurrence records submitted to GBIF during
255	the focal timeframe (Table 1), thus it was not possible to compute any statistics for them
256	either. All statistical analyses were performed using R statistical software (R Core Team
257	2019) with the following packages: base (Becker et al. 1988), MASS (Venables and Ripley
258	2002), stats (R Core Team 2019), regclass (Petrie 2020), and ggplot2 (Wickham 2016).

259 Findings

260 Taxonomic bias in the primary biodiversity data of South African animal taxa

261 More than 15 million occurrence records encompassing 49 South African animal classes

were submitted to the GBIF database from 1998 to 21 February 2020 (Table 1). These

263 occurrence records are from 88 datasets which are viewable on

doi.org/10.15468/dl.5upuwl. Among these datasets, the largest (with over 13 million

265 occurrence records) is from a volunteer/citizen science project called Southern African Bird

Atlas Project 2 which is published to GBIF by the Animal Demography Unit of the University

267 of Cape Town. The second largest dataset (with more than 1 million occurrence records) is

268 from historical bird ringing records (2005-2009) published by the South African National

269 Biodiversity Institute.

270

271 The number of GBIF occurrence records per class show strong bias towards Aves, while 272 some classes (e.g. Polychaeta and Branchiopoda) have comparatively fewer records (Table 273 1). A graphic representation of the ratio of GBIF occurrence records to the number of known 274 species per South African animal class (Fig. 1) illustrates the differences between most and 275 least represented classes, along with all other classes in between the two extremes. The 276 ratio of GBIF occurrence records to number of known species of birds is 18,584 while 277 Polychaeta has an average 0.01 occurrences submitted for each species (i.e. ratio of GBIF 278 occurrence records to number of known species = 0.01). The study sample has 23 279 underrepresented classes (with ratios lower than 1) and Polychaeta is the least represented 280 among those classes. Only 14 of the 49 classes submitted to GBIF have their ratio of 281 occurrence records to number of known species greater than one. The dashed section on 282 Fig. 1 represents the suggested lower target of 10 and upper target of 30 occurrence

records per species which would be ideal for various applications in biodiversity planning
and research, bearing in mind the difficulty of collecting occurrence data differs according to
species.

286

287 Insecta, the most species-rich South African animal class (SANBI Biodiversity Advisor 2020) 288 in the sample, accounts for 0.34% of the GBIF occurrence records under review here. The 289 second most species-rich class, Arachnida, accounts for 0.02% of the occurrence records in 290 the review sample. Aves accounts for 99.49% of occurrence records in the study sample yet 291 it is over 51 and six times less speciose than South African Insect and Arachnid species 292 respectively (SANBI Biodiversity Advisor 2020). The remaining 46 South African animal 293 classes jointly account for 0.15% of the total occurrence records submitted to GBIF between 294 1998 and 21 February 2020. Aves has the highest median number of occurrences per 295 species (419), while the rest of the classes in this sample have a median of 17 occurrence 296 records per species or less (Table 1). 297 298 Cultural salience and scientific interest as explanatory variables of taxonomic bias 299 The class with highest ratio of GBIF occurrence records to number of known species, Aves, 300 also has the highest cultural salience (Supplementary Material 2). Scientific interest is highest 301 for Insecta which has a lower ratio of occurrence records to known species in comparison to 302 Aves (1.25 vs 18,584). The least represented class in this study sample, Polychaeta, has 303 cultural salience and scientific interest in the mid-low ranges. Scaphopoda and Entognatha

304 have the lowest scientific interest and cultural salience respectively. Both classes are

305 underrepresented in GBIF occurrence records.

306

307	The correlation coefficient for GBIF occurrence records and cultural salience ($r = 0.799$)
308	indicates a strong positive linear relationship; the number of species occurrence records for
309	South African animals tend to increase with their cultural salience (Fig. 2) and this
310	relationship is statistically significant (p < 0.05). Scientific interest and GBIF occurrence
311	records show a weak linear relationship ($r = 0.011$) which is statistically non-significant (p =
312	0.95). Six outliers on each of the two plots were identified using the interquartile rule (Fig.
313	2). When these outliers are removed to achieve more evenly distributed data points,
314	cultural salience and GBIF occurrence records have a moderate positive relationship that is
315	statistically significant (r = 0.599, p < 0.05), while GBIF occurrence records and scientific
316	interest have a weak positive correlation that is also statistically significant (r = 0.396, p <
317	0.05).
318	
319	Higher resolution investigation of taxonomic bias using a subsample
319 320	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329
319320321	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21
319320321322	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets
 319 320 321 322 323 	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets (viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660
 319 320 321 322 323 324 	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets (viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660 occurrence records) is published by the South African Institute for Aquatic Biodiversity.
 319 320 321 322 323 324 325 	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets (viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660 occurrence records) is published by the South African Institute for Aquatic Biodiversity. Occurrence records of reptile species are from nine datasets (viewable on
 319 320 321 322 323 324 325 326 	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets (viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660 occurrence records) is published by the South African Institute for Aquatic Biodiversity. Occurrence records of reptile species are from nine datasets (viewable on doi.org/10.15468/dl.insbbc) and the largest among these had 946 occurrence records and is
 319 320 321 322 323 324 325 326 327 	Higher resolution investigation of taxonomic bias using a subsampleThe subsample used for higher resolution analysis of taxonomic bias consists of 4,329occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21February 2020 (Table 2). The amphibian occurrence records are from nine datasets(viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660occurrence records) is published by the South African Institute for Aquatic Biodiversity.Occurrence records of reptile species are from nine datasets (viewable ondoi.org/10.15468/dl.insbbc) and the largest among these had 946 occurrence records and ispublished by the South African National Biodiversity Institute. The GBIF occurrence records
 319 320 321 322 323 324 325 326 327 328 	Higher resolution investigation of taxonomic bias using a subsample The subsample used for higher resolution analysis of taxonomic bias consists of 4,329 occurrence records of South African herpetofauna submitted to GBIF between 1998 and 21 February 2020 (Table 2). The amphibian occurrence records are from nine datasets (viewable on doi.org/10.15468/dl.vletu9) and the largest of these datasets (with 660 occurrence records) is published by the South African Institute for Aquatic Biodiversity. Occurrence records of reptile species are from nine datasets (viewable on doi.org/10.15468/dl.insbbc) and the largest among these had 946 occurrence records and is published by the South African National Biodiversity Institute. The GBIF occurrence records of South African herpetofauna span 31 families, with Pipidae, Agamidae and Lamprophiidae

330	species is greater than one for 30 of the 31 reviewed families, and Amphisbaenidae is the
331	least represented.

333	Three families listed on the South African Animal Checklist (SANBI Biodiversity Advisor 2020)
334	do not have occurrence records submitted to GBIF during this study's timeframe (Table 2),
335	and are thus excluded from analysis. The most species-rich families (Pyxicephalidae,
336	Colubridae, Cordylidae, Gekkonidae, Scincidae) have median occurrence records per species
337	of five or lower. Pipidae and Pythonidae, which are among the least speciose of South
338	Africa's herpetofaunal families, have the highest median occurrence records per species
339	(Table 2). The GLM results (Table 3) show a positive and significant correlation between
340	GBIF occurrence records and cultural salience of the subsample. Correlations between the
341	subsample's GBIF occurrence records and scientific interest are non-significant (Table 3).
342	The GLM further shows that the interactions between cultural salience and scientific
343	interest have statistically significant and mostly negative influence on GBIF occurrence
344	records.
345	
346	Conservation status, cultural salience and scientific interest
347	No clear patterns emerge on the relationship between extinction risk of taxa, their cultural

348 salience and scientific interest (Table 4). Of the six IUCN taxonomic groupings of threatened

- 349 animals, mammals have the second highest cultural salience and scientific interest. They
- also have the second highest number of threatened species relative to the total number of
- 351 known species. Amphibians have the lowest cultural salience and scientific interest, and also
- have the highest number of threatened species relative to the number of known species.
- 353 Molluscs (which includes to Gastropoda, Bivalvia, and Cephalopoda as per IUCN taxonomic

- 354 grouping) have the third highest cultural salience and scientific interest, and the least
- 355 threatened species relative to the total number of known species in South Africa.

356 Interpretation and policy implications of findings

- 357 Taxonomic bias in the primary biodiversity data of South African animal taxa
- 358 This current study found biases in the primary biodiversity data of South African animal
- 359 taxa. There is a strong bias towards Aves in the South African context, and also at a global
- 360 scale (Troudet et al. 2017). Classes such as Polychaeta, Bivalvia and Arachnida are
- 361 underrepresented in both global (Troudet et al. 2017) and South African primary
- 362 biodiversity data (Fig. 1). Intrinsic factors may be contributing to the bias towards taxa such
- 363 as Aves and Mammalia since some are large (Zapponi et al. 2017), abundant (Royle and
- 364 Nichols 2003) and easily recognizable (Boakes et al. 2016). Arachnida and Insecta are both
- abundant and yet these two taxa are not as well represented as Aves and Mammalia.
- 366 Arachnid and Insect species are however smaller in size and thus more difficult to identify in
- 367 comparison to Aves and Mammalia. Herpetofauna are generally secretive with some cryptic
- 368 species which makes them difficult to identify, yet they are among the overrepresented taxa
- 369 within this study sample (Fig. 1). This reiterates an assertion made by Troudet et al. (2017)
- 370 that intrinsic features cannot solely account for the existing taxonomic bias. Approaches to
- 371 lessening taxonomic bias should take into consideration that extrinsic features also
- 372 contribute to underrepresentation of taxa in primary biodiversity data.
- 373
- 374 In addition to taxonomic bias, this review noted differences in class names and records
- among the biodiversity data sources consulted (Table 1). These discrepancies necessitate

376 standardisation across sources of biodiversity data (in this case GBIF and the South African 377 Animal Checklist) in order to maximise their value for biodiversity data applications. 378 Assessment of the taxonomic bias of 18 classes could not be completed as some of the data 379 required to compute the ratio of GBIF occurrence records to number of known species was 380 not available; 12 classes from the GBIF dataset did not have corresponding records on the 381 South African Animal Checklist (SANBI Biodiversity Advisor 2020), thus pointing out a need 382 for improved synchronisation as the institution producing this checklist also hosts the South 383 African node of GBIF. The remaining six of the unanalysed 18 classes do not have records 384 submitted to GBIF during the 22 year timeframe of this study. The lack of primary 385 biodiversity data records of six classes for this duration suggests lax monitoring of species 386 within those classes or the custodians of that data are not publishing it to GBIF. 387 388 The suggested lower and upper targets of species occurrence records do not seek to 389 introduce limits for what should constitute taxonomic bias. Hard limits are impractical since 390 species differ in population dynamics and primary biodiversity data requirements vary 391 according to application. Suggesting targets introduces a dimension that is often missing 392 from taxonomic bias research; the steps that follow confirmation of disparities in attention 393 received by taxa. If the disproportionate representation of taxa in biodiversity databases is 394 to be lessened, then guidelines of the target number of occurrence records per species 395 should be available. These guidelines should be cognisant of the multiple uses of primary 396 biodiversity data and intrinsic features that determine the ease with which species' 397 occurrences can be recorded. This lower target of 10 occurrence records per species 398 accommodates the secretive and less abundant taxa, and would serve to decrease under-399 representation of species. The upper target of 30 occurrence records per species would

400 guide data collection for all species and also serve to avoid perpetuating existing biases. If 401 the upper target per habitat was reached, then data collection resources could be used for 402 other species with lower occurrence records. Collecting biodiversity data beyond the upper 403 target would then be less urgent but still necessary as increased data are beneficial to the 404 accuracy of statistical models. The cumulative impacts of a minimum 10 occurrence records 405 per species per habitat is that average species occurrence records for each South African 406 animal class would end up being at least 10. In such a scenario, species that are easy to 407 detect would still have higher representation, but even the least represented species would 408 still have enough occurrence records to sufficiently inform research and planning. 409 410 Cultural salience and scientific interest as explanatory variables of taxonomic bias 411 Globally there is a mostly positive influence of public interests on representation of 412 taxonomic classes in primary biodiversity data and a few instances of positive correlation 413 between scientific interests and taxon representation (Troudet et al. 2017). A possible 414 contributor is the ease with which societal interest can be translated into primary 415 biodiversity data through citizen science platforms such as iNaturalist. The current study 416 found statistically significant and positive correlation between cultural salience and primary 417 biodiversity data of South African animal classes, while the correlation between scientific 418 interest and South African animal taxa representation is not statistically significant. The 419 regression results after removal of outliers with either high GBIF occurrence records, 420 cultural salience, or scientific interest (Fig. 2), suggest increased influence of scientific 421 interest on species occurrence records when there is less biased focus on taxa. 422

Environmental Reviews

423	Higher resolution investigation of taxonomic bias using a subsample
424	A higher resolution analysis of taxonomic bias found positive and statistically significant
425	correlation between cultural salience and GBIF occurrence records of a subsample of
426	herpetofaunal families. No statistically significant correlation was found between GBIF
427	occurrence records and scientific interest in this subsample. Scientific interest's effect on
428	primary biodiversity data accumulation appears negligible but should not be overlooked as
429	its interaction with cultural salience has a statistically significant influence on GBIF
430	occurrence records (Table 3). Statistically significant interaction effects indicate that there is
431	possibly an interaction between cultural salience and scientific interest that is affecting the
432	number of GBIF occurrence records per South African animal taxon. Thus, the influence of
433	cultural salience on GBIF occurrence per species may depend on the scientific interest a
434	species has (and vice versa). Consequently, it may not be possible to solely attribute
435	taxonomic bias in GBIF occurrence records to cultural salience as these primary biodiversity
436	data are collected by both scientists and non-scientists. There is also evidence that societal
437	and scientific interests in biodiversity matters are not mutually exclusive (Eisner et al. 1995;
438	Wilson et al. 2007; Martín-López, et al 2009).

439

Herpetofauna being generally well represented in GBIF occurrence records at both national
and international scales is good for herpetofaunal research and conservation as it means
there are generally sufficient amounts of primary biodiversity data to work with.
Furthermore, increased representation of herpetofauna in primary biodiversity data is an
encouraging result considering that they are generally understudied (Christoffel and Lepczyk

445 2012), the public mostly has negative perceptions about them (Reimer et al. 2013; Tarrant

446 et al. 2016), and their global populations are declining (McCallum 2007).

448 Conservation status, cultural salience and scientific interest

449 This study's inferences about the relationship between extinction risk, cultural salience and 450 scientific interest are based on data with a small scale of focus and further investigation at a 451 higher resolution (i.e. species level) is required to better understand the relationship. 452 Published literature shows mixed results with regard to the correlation of extinction risk 453 with cultural salience and scientific interest. Higher resolution investigations from previous 454 research found that some taxa with high societal preference also have a high extinction risk 455 (Courchamp et al. 2018). Society's taxonomic preferences result from various reasons 456 (including species abundance and charisma) and are not necessarily motivated by concern 457 for a taxon's welfare (Davies et al. 2019). For endangered taxa that receive increased 458 attention, the focus is biased towards taxa that fit certain criteria. Threatened vertebrates 459 receive greater scientific interest than threatened invertebrates (Donaldson et al. 2016). 460 Furthermore, the number of IUCN threat status assessments for invertebrates is much lower 461 than vertebrate assessments (Eisenhauer et al. 2019). Threatened vertebrates have higher 462 cultural salience if they are birds or mammals (Davies et al. 2019). Species that are 463 threatened generally obtain higher scientific interest than non-threatened species and this 464 is due to conservation efforts being based on IUCN's threatened species categories (Martín-465 López et al. 2011).

466

Previous research has shown that high cultural salience alone does not correlate with
increased species protection (Courchamp et al. 2018), however, higher cultural salience
alone can drive funds towards species protection (Simberloff 1998). These funds can in turn
be used to increase scientific focus on underfunded biodiversity research necessary to

471 inform conservation policy. Cultural salience influences scientific interest (Wilson et al. 472 2007) and society plays a role in biodiversity research and planning (Martín-López et al 473 2009). There are also suggestions that biodiversity protection is most effective when based 474 on scientific knowledge and has societal approval (Eisner et al. 1995). Biodiversity 475 researchers and managers often direct scientific interest and cultural salience towards 476 certain species for the benefit of many other species. This surrogate species concept 477 chooses species to be proxies for ecosystems (Favreau et al. 2006) or conservation problems 478 (Dietz et al. 1994). By updating the surrogate species framework to incorporate the 479 suggested targets of 10 - 30 occurrence records per species, its scope of benefits can be 480 broadened to include lessening of persisting taxonomic bias. In this way the outcomes of 481 dedicating resources towards surrogates would include 10 - 30 occurrence records per 482 surrogate species and the species meant to benefit from protection of the surrogate.

Concluding Remarks and Recommendations

483

484	This study quantifies taxonomic bias in the primary biodiversity data of animal taxa in a
485	megadiverse country and shows there is a severe bias towards birds while some classes are
486	underrepresented. Statistical analysis suggests that cultural salience has greater influence
487	on the noted taxonomic bias than scientific interest. A high resolution analysis of taxonomic
488	bias with a subsample, also shows cultural salience to have stronger influence and
489	additionally suggests there is statistical interaction between the two explanatory variables.
490	No clear correlation was found in the relationship between a taxon's extinction risk, cultural
491	salience and scientific interest.
492	
493	The intrinsic traits of species along with the limitations of ecological research make it
494	difficult to completely avoid taxonomic bias. It is more feasible to investigate extrinsic
495	factors (such as cultural salience and scientific interest) and use the findings to avoid the
496	current situation where taxonomic bias in primary biodiversity data has prevailed for
497	decades and the majority of animal taxa of a megadiverse country are underrepresented.
498	Approaches that will, at the least, increase representation of severely underrepresented
499	taxa in order to lessen persistent taxonomic bias are required. Once taxonomic bias has
500	been quantified, additional research time should be dedicated to finding ways to address
501	persisting biases. Our recommendation to introduce soft targets of between 10 and 30
502	species occurrence records per habitat seeks to increase representation of
503	underrepresented taxa in primary biodiversity data. The species occurrence targets can also
504	be incorporated into surrogate species frameworks for their benefits to be extended to
505	include the lessening of persisting taxonomic bias. Investigations of extrinsic factors should
506	also consider how interactions between the explanatory variables may influence taxonomic

- 507 bias. Given the possibility of interaction effects between societal and scientific preferences,
- 508 it will be worth researching the extent of this interaction between explanatory variables and
- 509 how it can be used to lessen taxonomic bias.

510 Acknowledgements

- 511 This research is made possible by a bilateral scientific cooperation between North-West
- 512 University and Hasselt University. Financial support for FMP was provided by the National
- 513 Research Foundation (UID: 114663), North-West University, and the Flemish Interuniversity
- 514 Council (VLIR-UOS) Global Minds program (Contract Number: R-9363). MPMV is supported
- 515 by the Special Research Fund of Hasselt University (BOF20TT06). No conflict of interest to be
- 516 declared.

517	References
518	Ball-Damerow, J.E., Brenskelle, L., Barve, N., Soltis, PS., Sierwald, P., Bieler, R., LaFrance, R.,
519	Ariño, A.H., and Guralnick, R.P. 2019. Research applications of primary biodiversity
520	databases in the digital age [online]. PloS ONE. 14 (9): e0215794.
521	doi.org/10.1371/journal.pone.0215794
522	
523	Becker, R.A., Chambers, J.M., and Wilks, A.R. 1988. A Programming Environment for Data
524	Analysis and Graphics (Wadsworth & Brooks/Cole computer science series). Taylor and
525	Francis, Abingdon-on-Thames.
526	
527	Boakes, E.H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D.B., and Haklay, M. 2016.
528	Patterns of contribution to citizen science biodiversity projects increase understanding of
529	volunteers' recording behaviour. Sci. Rep. 6 : 1–11.
530	
531	Boyd, D., and Crawford, K. 2012. Critical questions for big data. Inform. Commun. Soc. 15:
532	662–679.
533	
534	Christoffel, R.A., and Lepczyk, C.A. 2012. Representation of herpetofauna in wildlife research
535	journals. J. Wildl. Manage. 76 (4): 661–669.
536	
537	Clark, J.A., and May, R.M. 2002. How biased are we?: Even now, conservation research is
538	still lopsided. Conserv. Practice. 3 (3): 28–29.
539	

540	Cohen, J., and Cohen, L. 1995. Statistics for ornithologists. 2nd Guide. British Trust for
541	Ornithology, Thetford.
542	
543	Correia, R.A., Jepson, P.R., Malhado, A.C.M., and Ladle, R.J. 2016 Familiarity breeds content:
544	assessing bird species popularity with culturomics [online]. Peerj. 4: e1728.
545	doi.org/10.7717/peerj.1728
546	
547	Correia, R.A., Jepson, P., Malhado, A.C. and Ladle, R.J. 2017. Internet scientific name
548	frequency as an indicator of cultural salience of biodiversity. Ecol. Indic. 78 : 549–555.
549	
550	Courchamp, F., Jarić, I., Albert, C., Meinard, Y., Ripple, W.J., and Chapron, G. 2018. The
551	paradoxical extinction of the most charismatic animals [online]. PLoS. Biol. 16 (4): e2003997.
552	doi.org/10.1371/journal.pbio.2003997
553	
554	Davies, T., Cowley, A., Bennie, J., Leyshon, C., Inger, R., Carter, H., Robinson, B., Duffy, J.,
555	Casalegno, S., Lambert, G., and Gaston, K. 2019. Popular interest in vertebrates does not
556	reflect extinction risk and is associated with bias in conservation investment [online]. PloS
557	ONE. 14 (2): e0212101. doi.org/10.1371/journal.pone.0212101
558	
559	Dietz, J.M., Dietz, A.L., and Nagagata E.Y. 1994. The effective use of flagship species for
560	conservation of biodiversity: the example of lion tamarins in Brazil. <i>In</i> Creative conservation:
561	interactive management of wild and captive animals. <i>Edited by</i> P.J.S. Olney, G.M. Mace and
562	A.T.C. Feistner. Chapman and Hall, London. pp. 32–49.
563	

564	Di Marco, M., Chapman, S., Althor, G., Kearney, S., Besancon, C., Butt, N., Maina, J.M.,
565	Possingham, H.P., von Bieberstein, K.R., Venter, O., and Watson, J.E. 2017. Changing trends
566	and persisting biases in three decades of conservation science. Glob. Ecol. Conserv. 10: 32–
567	42.
568	
569	Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J., and Kerr, J.T.
570	2016. Taxonomic bias and international biodiversity conservation research. FACETS. 1: 105–
571	113. doi:10.1139/facets-2016-0011
572	
573	Ducarme, F., Luque, G.M., and Courchamp, F. 2013. What are "charismatic species" for
574	conservation biologists. BioSciences Master Reviews. 10 : 1–8.
575	
576	Edwards, J.L. 2004. Research and societal benefits of the Global Biodiversity Information
577	Facility. BioScience. 54 (6): 485–486.
578	
579	Eisenhauer, N., Bonn, A., and Guerra, C.A. 2019. Recognizing the quiet extinction of
580	invertebrates. Nat. Commun. 10(50). <u>https://doi.org/10.1038/s41467-018-07916-1</u>
581	
582	Eisner, T., Lubchenco, J., Wilson, E.O., Wilcove, D.S., and Bean, M.J. 1995. Building a
583	scientifically sound policy for protecting endangered species. Science. 269(5228): 1231–
584	1233.
505	

586	Favreau, J.M., Drew C.A., Hess G.R., Rubino M.J., Koch F.H., and Eschelbach K.A. 2006.
587	Recommendations for assessing the effectiveness of surrogate species approaches.
588	Biodivers. Conserv. 15 :3949–3969.
589	
590	Feeley, K., Stroud, J., and Perez, T. 2016. Most 'global' reviews of species' responses to
591	climate change are not truly global. Divers. Distrib. 23 :231–234.
592	
593	Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., and Neufeld, D. 2007. A web-based GIS
594	tool for exploring the world's biodiversity: The Global Biodiversity Information Facility
595	Mapping and Analysis Portal Application (GBIF-MAPA). Ecol. inform. 2 (1): 49–60.
596	
597	Funk, S.M., and Rusowsky, D. 2014. The importance of cultural knowledge and scale for
598	analysing internet search data as a proxy for public interest toward the environment.
599	Biodivers. Conserv. 23:3101–3112. <u>https://doi.org/10.1007/s10531-014-0767-6</u>
600	
601	GBIF.org. 2020a. Become a publisher [online]. Available from:
602	https://www.gbif.org/become-a-publisher [accessed 12 December 2020].
603	
604	GBIF.org. 2020b. The GBIF Network [online]. Available from: <u>https://www.gbif.org</u> [accessed
605	12 December 2020].
606	
607	Gordon, E.R., Butt, N., Rosner-Katz, H., Binley, A.D., and Bennett, J.R. 2020. Relative costs of
608	conserving threatened species across taxonomic groups. Conserv. Biol. 34 (1): 276–281.
609	

610	Griffiths, R.A., and Dos Santos, M. 2012. Trends in conservation biology: Progress or
611	procrastination in a new millennium?. Biol. Conserv. 153 : 153–158.
612	
613	Harrell Jr., F.E. 2001. Regression strategies. Springer-Verlag, New York, NY.
614	
615	Hernandez, P.A., Graham, C.H., Master, L.L., and Albert, D.L. 2006. The effect of sample size
616	and species characteristics on performance of different species distribution modeling
617	methods. Ecography 29 : 773–785.
618	
619	Huang, X., Hawkins, B.A., and Qiao, G. 2013. Biodiversity data sharing: Will peer-reviewed
620	data papers work?. BioScience 63(1): 5–6.
621	
622	IUCN (International Union for Conservation of Nature). 2020. IUCN Red List version 2020-1:
623	Table 5 [online]. Available from https://www.iucnredlist.org [accessed 21 February 2020].
624	
625	Jarić, I., Courchamp, F., Gessner, J., and Roberts, D.L. 2016. Data mining in conservation
626	research using Latin and vernacular species names. PeerJ. 4: e2202. doi:10.7717/peerj.2202
627	
628	Kim, J.Y., Do, Y., Im, R.Y., Kim, G.Y., and Joo, G.J. 2014. Use of large web-based data to
629	identify public interest and trends related to endangered species. Biodivers. Conserv.
630	23 (12): 2961–2984.
631	

- 632 Lira-Noriega, A., Soberón, J., Navarro-Sigüenza, A.G., Nakazawa, Y., and Peterson, A.T. 2007.
- 633 Scale dependency of diversity components estimated from primary biodiversity data and
- 634 distribution maps. Divers. Distrib. **13**(2): 185–195.
- 635
- 636 McCallum, M.L. 2007. Amphibian extinction or decline? Current declines dwarf background
- 637 extinction. J. Herpetol. **41**: 483–491.

- 639 McKinney, M. 1999. High rates of extinction and threat in poorly studied taxa. Conserv. Biol.
- 640 **13**: 1273–1281.

641

- 642 Martín-López, B., González, J.A., and Montes, C. 2011. The pitfall-trap of species
- 643 conservation priority setting. Biodiver. Conserv. **20**(3): 663–82.

644

- 645 Martín-López, B., Montes, C., Ramírez, L., and Benayas, J. 2009. What drives policy decision-
- 646 making related to species conservation? Biol. Conserv. **142**: 1370–1380.

647

- 648 Mittermeier, R.A. 1988. Primate diversity and the tropical forest: case studies
- 649 from Brazil and Madagascar and the importance of the megadiversity countries. In
- 650 Biodiversity. Edited by E.O. Wilson. National Academies Press, Washington, DC. pp. 145-

651 154.

652

- 653 Mittermeier, R.A., Gil, P.R., and Mittermeier, C.G. 1997. Megadiversity: Earth's Biologically
- 654 Wealthiest Nations. Conservation International, Arlington.

656	National Environmental Management Act see South Africa.
657	
658	Pawar, S. 2003. Taxonomic chauvinism and the methodologically challenged. BioScience 53
659	861.
660	
661	Petrie, A. 2020. regclass: Tools for an Introductory Class in Regression and Modeling. R
662	package version 1.6. Available from https://CRAN.R-project.org/package=regclass
663	
664	Ponder, W.F. 1992. Bias and biodiversity. Aust. Zool. 28 (1-4): 47–51.
665	
666	R Core Team. 2019. R: A language and environment for statistical computing. R Foundation
667	for Statistical Computing, Vienna, Austria. Available from <u>https://www.R-project.org/</u>
668	
669	Reimer, A., Mase, A., Mulvaney, K., Mullendore, N., Perry-Hill, R., and Prokopy, L. 2013. The
670	impact of information and familiarity on public attitudes toward the eastern hellbender.
671	Anim. Conserv. 17 : 235–243.
672	
673	Royle, J.A., and Nichols, J.D. 2003. Estimating abundance from repeated presence–absence
674	data or point counts. Ecology. 84 : 777–790.
675	
676	Russell, L.M. 1984. The Fauna of India and the Adjacent Countries, Homoptera: Aphidoidea.
677	Bulletin of the ESA. 30 (2): 56.
678	

- 679 SANBI Biodiversity Advisor (South African National Biodiversity Institute Biodiversity
- 680 Advisor). 2020. South African Animal Checklist [online]. Available from
- 681 http://biodiversityadvisor.sanbi.org/research-and-modelling/checklists-and-encyclopaedia-
- 682 of-life/south-african-animal-checklist/). [accessed 21 February 2020].
- 683
- 684 Schuetz, J., Soykan, C.U., Distler, T., and Langham, G. 2015. Searching for backyard birds in
- 685 virtual worlds: Internet queries mirror real species distributions. Biodivers. Conserv. 24:
- 686 1147–1154. https://doi.org/10.1007/s10531-014-0847-7
- 687
- 688 Seddon, P., Soorae, P., and Launay, F. 2005. Taxonomic bias in reintroduction projects.
- 689 Anim. Conserv. 8: 51–58.
- 690
- 691 Simberloff, D. 1998. Flagships, umbrellas, and keystones: is single-species management
- 692 passe in the landscape era? Biol. Conserv **83**: 247–257.
- 693
- 694 Soberón, J., and Peterson, T. 2004. Biodiversity informatics: managing and applying primary
- 695 biodiversity data. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **359**(1444): 689–698.
- 696
- 697 South Africa. 1998. National Environmental Management Act 107 of 1998.
- 698
- 699 Stahlschmidt, Z.R. 2011. Taxonomic chauvinism revisited: insight from parental care
- 700 research. Plos ONE. 6(8): e24192. doi.org/10.1371/journal.pone.0024192
- 701

702	Tarrant, J., Kruger, D., and Du Preez, L.H. 2016. Do public attitudes affect conservation
703	effort? Using a questionnaire-based survey to assess perceptions, beliefs and superstitions
704	associated with frogs in South Africa. Afr. Zool. 51 (1): 3–20.
705	
706	Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., and Legendre, F. 2017. Taxonomic bias
707	in biodiversity data and societal preferences. Sci. Rep. 7(1): 9132.
708	
709	Troumbis, A.Y. 2017. Google Trends and cycles of public interest in biodiversity: the animal
710	spirits effect. Biodivers. Conserv. 26, 3421–3443. https://doi.org/10.1007/s10531-017-1413-
711	x
712	
713	Van Proosdij, A.S., Sosef, M.S., Wieringa, J.J., and Raes, N. 2016. Minimum required number
714	of specimen records to develop accurate species distribution models. Ecography, 39 (6):
715	542–552.
716	
717	Venables, W.N., and Ripley, B.D. 2002. Modern Applied Statistics with S. 4th ed. Springer,
718	New York, NY. Available from http://www.stats.ox.ac.uk/pub/MASS4
719	
720	Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
721	
722	Wilson, J., Procheş, Ş., Braschler, B., Dixon, E., and Richardson, D. 2007. The (bio)diversity of
723	science reflects the interests of society. Front. Ecol. Environ. 5 : 409–414
724	

- 725 Zapponi, L., Cini, A., Bardiani, M., Hardersen, S., Maura, M., Maurizi, E., De Zan, L.R., Audisio,
- 726 P., Bologna, M.A., Carpaneto, G.M., and Roversi, P.F. 2017. Citizen science data as an
- 727 efficient tool for mapping protected saproxylic beetles. Biol. Conser. **208**: 139–145.

728 Tables

Table 1: Occurrence data statistics for South African animal taxa arranged alphabetically by
class name. The occurrence records obtained from GBIF (2020) are for the period from 1998
to 21 February 2020, and exclude fossil records and occurrences designated as unknown.
Total number of known species is obtained from the South African Animal Checklist (SANBI

733 Biodiversity Advisor 2020).

Class*	GBIF	Total	Median: GBIF	Median	Occurrence
	occurrence	known	occurrence	absolute	records to
	records	species	records	deviation	known species
					ratio
Actinopterygii	14,240	2,200	3	2	6.47
Adenophorea ^a	5		2	2	
Amphibia	1,148	123	4	2	9.33
Anthozoa	534	174	4	3	3.07
Aplacophora ^b		2			
Appendicularia ^b		80			
Arachnida	2,925	6,630	2	1	0.44
Ascidiacea	559	176	17	13	3.18
Asteroidea	54	91	1	0	0.59
Aves	15,870,476	854	419	418	18,584
Bivalvia	127	650	1	0	0.20
Branchiopoda	9	120	1	0	0.08
Cephalochordata ^b		1			

Cephalopoda	18	195	2	0	0.09
Cestoda	32	83	1	0	0.39
Chilopoda	147	141	2	1	1.04
Chondrichthyes ^c	549	188	2	1	2.92
Clitellata	78	102	2	1	0.76
Crinoidea	30	19	11	6	1.58
Cubozoa	1	2	1	0	0.50
Demospongiae ^a	409		10	6	
Diplopoda	209	462	8	6	0.45
Echinoidea	18	59	2	0	0.31
Entognatha	10	195	2	2	0.05
Eoacanthocephala ^a	1		1	0	
Eurotatoria ^a	1		1	0	
Gastropoda	553	2,262	1	0	0.24
Gordioida ^a	2		2	0	
Gymnolaemata ^a	440		4	4	
Hexanauplia ^a	3		1	0	
Holothuroidea	30	122	1	0	0.25
Hydrozoa	203	457	4	3	0.44
Insecta	54,681	43,893	3	2	1.25
Malacostraca	202	1,763	2	2	0.11
Mammalia	617	307	2	1	2.01
Maxillopoda	12	511	1	0	0.02

Monogenea	9	49	2	0	0.18
Myxini	7	4	2	1	1.75
Ophiuroidea	128	119	2	1	1.08
Ostracoda	18	165	1	0	0.11
Pauropoda ^b		2			
Phylactolaemata ^a	17		2	1	
Polychaeta	9	760	1	0	0.01
Polyplacophora	7	29	1	0	0.24
Pycnogonida	5	101	5	0	0.05
Reptilia	3,181	381	4	3	8.35
Rhynchonellata ^a	30		15	14	
Sarcopterygii	17	3	17	0	5.67
Scaphopoda	2	16	1	0	0.13
Scyphozoa ^b		10			
Secernentea ^a	166		1	0	
Stenolaemata ^a	55		1	0	
Symphyla ^a	1		1	0	
Trematoda	58	72	2	1	0.81
Turbellaria ^b		42			
Unassigned ^d	770	-	-	-	-
Total	15,952,803	63,615	578	497	18,638.15

* Taxon names copied verbatim from GBIF (2020) and South African Animal

Checklist (SANBI Biodiversity Advisor 2020), some classes on this list are paraphyletic.

 a Taxa not listed on South African Animal Checklist (SANBI Biodiversity Advisor

2020).

^b No records submitted to GBIF during this study's period of interest.

^c Listed as two separate classes on GBIF, namely Elasmobranchii and Holocephali.

^d Animal occurrence records that were not assigned to any class or lower taxonomic

level

Table 2: Occurrence data statistics for South Africa's herpetofaunal taxa. The table is
arranged alphabetically by class, then family. These occurrence records are obtained from
GBIF (2020) for the period from 1998 to 21 February 2020, and exclude fossil records and
occurrences designated as unknown. The total number of known species is obtained from
the South African Animal Checklist (SANBI Biodiversity Advisor 2020).

Family*	GBIF	Total	Median:	Median	Occurrence
	occurrence	known	GBIF	absolute	records to
	records	species	occurrence	deviation	known
			records		species ratio
Amphibians					
Arthroleptidae	7	5	3	0	1.40
Brevicipitidae	62	15	2	1	4.13
Bufonidae	168	17	5	3	9.88
Heleophrynidae	13	7	2	1	1.86
Hemisotidae	3	3	2	0	1.00
Hyperoliidae	63	18	4	1	3.50
Microhylidae	15	2	8	2	7.50
Phrynobatrachidae	15	3	6	4	3.00
Pipidae	221	3	11	9	73.67
Ptychadenidae	15	10	2	2	1.50
Pyxicephalidae	559	39	4	3	14.33
Rhacophoridae	7	1	7	0	7.00
Total	1,148	123	56	26	128.16

Reptiles					
Agamidae	182	7	7	2	26.00
Amphisbaenidae	9	10	2	1	0.90
Chamaeleonidae	103	21	8	7	4.90
Cheloniidae ^a	0	4	0		
Colubridae	179	68	2	2	2.63
Cordylidae	184	44	2	2	4.18
Crocodylidae	1	1	1	0	1.00
Dermochelyidae ^a	0	1			
Elapidae	70	15	2	2	4.67
Gekkonidae	856	71	5	4	12.06
Gerrhosauridae	80	13	3	2	6.15
Hydrophiidae ^a	0	1			
Lacertidae	157	25	5	3	6.28
Lamprophiidae	407	12	5	3	33.92
Leptotyphlopidae	46	8	10	8	5.75
Pelomedusidae	18	5	9	2	3.60
Pythonidae	11	1	11	0	11.00
Scincidae	556	65	5	4	8.55
Testudinidae	159	15	2	1	10.60
Typhlopidae	45	9	1	0	5.00
Varanidae	24	2	10	3	12.00
Viperidae	94	13	8	6	7.23

Total 3,181	411	98	52	166.42
-------------	-----	----	----	--------

* Family names copied verbatim from GBIF (2020) and South African Animal

Checklist (SANBI Biodiversity Advisor 2020).

^{*a*} No records submitted to GBIF for this study's focal timespan.

|--|

- variable (GBIF occurrence records of herpetofaunal families) and two explanatory variables
- 741 (cultural salience and scientific interest) and their combined influence. (*) indicates a
- 742 significant p-value at 5% threshold. (+) and (-) respectively indicate positive and negative
- 743 correlation of explanatory variable with number of GBIF occurrence records.

Herpetofaunal family ^a	Cultural salience	Scientific interest	Interaction
	influence p-value	influence p-value	influence p-value
Brevicipitidae	(+) 0.002*	(+) 1.356	(-) 0.004*
Bufonidae	(+) 0.003*	(+) 0.093	(-) 0.000*
Colubridae	(+) 0.002*	(-) 0.549	(-) 0.000*
Cordylidae	(+) 0.004*	(+) 0.298	(-) 0.000*
Elapidae	(+) 0.001*	(+) 0.245	(-) 0.000*
Gekkonidae	(+) 0.001*	(+) 0.463	(+) 0.000*
Lacertidae	(+) 0.003*	(+) 0.537	(-) 0.001*
Lamprophiidae	(+) 0.001*	(+) 1.122	(-) 0.001*
Pyxicephalidae	(+) 0.002*	(+) 0.834	(-) 0.001*
Scincidae	(+) 0.001*	(-) 0.165	(-) 0.001*
Testudinidae	(+) 0.002*	(+) 0.079	(-) 0.000*

^{*a*} Family names copied verbatim from GBIF (2020) and South African Animal

Checklist (SANBI Biodiversity Advisor 2020).

- 744 **Table 4:** Comparison of extinction risk with cultural salience and scientific interest of South
- 745 African animal taxa. Columns arranged by ascending order of the magnitude of cultural
- 746 salience and scientific interest out of the 49 classes under review here.

Cultural salience	IUCN	Threatened	Known species ⁺
ranking*	Taxonomic	Species	
	grouping		
1	Birds	54	854
2	Mammals	30	307
3	Molluscs ^a	22	3107
4	Fishes ^{b,c}	128	2395
6	Reptiles ^b	19	381
8	Amphibians	16	132
Scientific interest	IUCN	Threatened	Known species ⁺
ranking*	Taxonomic	Species	
ranking*	Taxonomic grouping	Species	
2	Taxonomic grouping Fishes ^{b,c}	Species	2395
2 4	Taxonomic grouping Fishes ^{b,c} Mammals	Species 128 30	2395 307
2 4 5	TaxonomicgroupingFishes ^{b,c} MammalsMolluscs ^a	Species 128 30 22	2395 307 3107
2 4 5 8	TaxonomicgroupingFishes ^{b,c} MammalsMolluscs ^a Reptiles ^b	Species 128 30 22 19	2395 307 3107 381
2 4 5 8 11	TaxonomicgroupingFishes ^{b,c} MammalsMolluscs ^a Reptiles ^b Birds	Species 128 30 22 19 54	2395 307 3107 381 854

* Ranking out of the 49 classes in this study's sample.

⁺ Number of known species obtained from the South African Animal Checklist (SANBI Biodiversity Advisor 2020).

^a Molluscs collectively refers to Gastropoda, Bivalvia, Cephalopoda.
 ^b The conservation status of these taxonomic groupings has not
 been fully assessed. Numbers of threatened species should be
 interpreted as the number of species known to be threatened
 within those species that have been assessed, and not as the overall
 number of threatened species within a grouping (IUCN 2020).
 ^c Fishes collectively refers to Actinopterygii, Chondrichthyes, Myxini,
 and Sarcopterygii.

747	Figure Captions
748	Fig. 1: Taxonomic bias in South African animal classes. The number '1' on the horizontal axis
749	represents the point where occurrence records are equal to total number of species per
750	class. Ordering is by decreasing ratio of representation where, value >1 denotes over-
751	representation and value <1 denotes underrepresentation of a class in GBIF occurrence
752	records. Log transformation is used on the horizontal axis due to disparities between least
753	and most represented species. The dashed section suggests a 10-30 occurrence per species
754	threshold that would be ideal for various biodiversity data uses. Taxa with no records
755	submitted to GBIF (2020) for this study's timespan or not listed on the South African Animal
756	Checklist (SANBI Biodiversity Advisor 2020) are omitted from this graph as it was not
757	possible to calculate their occurrence records to known species ratio.
758	
759	Fig. 2: Correlations of GBIF occurrence records with cultural salience and scientific interest
760	of South African Animal taxa. The occurrence records of South African taxa obtained from
761	GBIF (2020) are plotted on the vertical axis. Cultural salience is represented by frequency of
762	taxon names in web pages, and scientific interest is represented by number of scientific
763	articles focused on a taxon are plotted on the horizontal axis. Axes are Log transformed due

to disparities among the plotted variables. Circled points represent outliers.



Fig. 1 Taxonomic bias in South African animal classes. The number '1' on the horizontal axis represents the point where occurrence records are equal to total number of species per class. Ordering is by decreasing ratio of representation where, value >1 denotes over-representation and value <1 denotes underrepresentation of a class in GBIF occurrence records. Log transformation is used on the horizontal axis due to disparities between least and most represented species. The dashed section suggests a 10-30 occurrence per species threshold that would be ideal for various biodiversity data uses. Taxa with no records submitted to GBIF (2020) for this study's timespan or not listed on the South African Animal Checklist (SANBI Biodiversity Advisor 2020) are omitted from this graph as it was not possible to calculate their occurrence records to known species ratio.



Fig. 2: Correlations of GBIF occurrence records with cultural salience and scientific interest of South African Animal taxa. The occurrence records of South African taxa obtained from GBIF (2020) are plotted on the vertical axis. Cultural salience is represented by frequency of taxon names in web pages, and scientific interest is represented by number of scientific articles focused on a taxon are plotted on the horizontal axis. Axes are Log transformed due to disparities among the plotted variables. Circled points represent outliers.