






# Sampling bias in reptile occurrence data for the Kruger National Park



## Authors:

Jody M. Barends<sup>1</sup>   
 Darren W. Pietersen<sup>2</sup>   
 Guinevere Zambatis<sup>3</sup>   
 Donovan R.C. Tye<sup>4</sup>   
 Bryan Maritz<sup>1</sup> 

## Affiliations:

<sup>1</sup>Department of Biodiversity and Conservation Biology, University of the Western Cape, Cape Town, South Africa

<sup>2</sup>Department of Zoology and Entomology, University of Pretoria, Pretoria, South Africa

<sup>3</sup>Scientific Services, Kruger National Park, Skukuza, South Africa

<sup>4</sup>Organization for Tropical Studies, Kruger National Park, Skukuza, South Africa

## Corresponding author:

Jody Barends,  
 jbarends99@gmail.com

## Dates:

Received: 25 July 2019

Accepted: 24 Feb. 2020

Published: 11 May 2020

## How to cite this article:

Barends, J.M., Pietersen, D.W., Zambatis, G., Tye, D.R.C. & Maritz, B., 2020, 'Sampling bias in reptile occurrence data for the Kruger National Park', *Koedoe* 62(1), a1579. <https://doi.org/10.4102/koedoe.v62i1.1579>

## Copyright:

© 2020. The Authors.  
 Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

## Read online:



Scan this QR code with your smart phone or mobile device to read online.

To effectively conserve and manage species, it is important to (1) understand how they are spatially distributed across the globe at both broad and fine spatial resolutions and (2) elucidate the determinants of these distributions. However, information pertaining to the distributions of many species remains poor as occurrence data are often scarce or collected with varying motivations, making the resulting patterns susceptible to sampling bias. Exacerbating an already limited quantity of occurrence data with an assortment of biases hinders their effectiveness for research, thus making it important to identify and understand the biases present within species occurrence data sets. We quantitatively assessed occurrence records of 126 reptile species occurring in the Kruger National Park (KNP), South Africa, to quantify the severity of sampling bias within this data set. We collated a data set of 7118 occurrence records from museum, literature and citizen science sources and analysed these at a biologically relevant spatial resolution of 1 km × 1 km. As a result of logistical challenges associated with sampling in KNP, approximately 92% of KNP is data deficient for reptile occurrences at the 1 km × 1 km resolution. Additionally, the spatial coverage of available occurrences varied at species and family levels, and the majority of occurrence records were strongly associated with publicly accessible human infrastructure. Furthermore, we found that sampled areas within KNP were not necessarily ecologically representative of KNP as a whole, suggesting that areas of unique environmental space remain to be sampled. Our findings highlight the need for substantially greater sampling effort for reptiles across KNP and emphasise the need to carefully consider the sampling biases within existing data should these be used for conservation management decision-making. Modelling species distributions could potentially serve as a short-term solution, but a concomitant increase in surveys across the park is needed.

**Conservation implications:** The sampling biases present within KNP reptile occurrence data inhibit the inference of fine-scale species distributions within and across the park, which limits the usage of these data towards meaningfully informing conservation management decisions as applicable to reptile species in KNP.

**Keywords:** conservation management; Kruger National Park; occurrence data; reptile fauna; sampling bias.

## Introduction

Effective conservation and management of organisms require an understanding of how species are spatially distributed at both broad and fine spatial resolutions, and ideally also the underlying determinants of their distribution patterns (Hurlbert & Jetz 2007; Kery 2011). However, species geographic data that may help inform conservation management decisions are often limited and biased in their collection strategies (Franklin 2010). For example, although museum databases often include occurrence data of collected specimens, the principal purpose of most museum collections is to act as reference catalogues for species identification rather than for species distribution mapping (Newbold 2010). It is important to note that although several museum specimens are collected directly as a result of systematic sampling, many specimens are collected opportunistically (Kadmon, Farber & Danin 2004; Pyke & Ehrlich 2010). As a result, collection effort and spatial coverage within museum data naturally vary depending on the interests of the collection. Despite this, a recently increased urgency in the need for species distribution information has placed a greater emphasis on the use of museum databases for amassing species occurrence records (Syfert, Smith & Coomes 2013).

In recent years, the capture of museum data within electronic databases, the establishment and continued activities of atlas projects, and the growth of citizen science projects have provided a wealth of species occurrence data that are accessible online (Newbold 2010). These data are

undoubtedly valuable but are subject to a multitude of biases, errors and uncertainties that need to be considered should these data be used for environmental research. Generally, species occurrence records are susceptible to geospatial or taxonomic sampling biases and on their own do not explain the full extents of species distributions (Bird et al. 2014; Botts, Erasmus & Alexander 2011; Reddy & Davalos 2003). For example, museum data are often biased towards heavily sampled areas (Newbold 2010), atlas data tend to be vulnerable to omission errors (Botts et al. 2011) and citizen science records are often imprecise (Geldmann et al. 2016; McGrath et al. 2015).

For rare and understudied species, bias in occurrence data sets exacerbates an already severe issue of misinformation and overall data deficiency (Reddy & Davalos 2003). With recorded occurrences of these animals already limited, the presence of an assortment of sampling biases within databases further restricts our understanding of these species' distributions and curtails our ability to manage them effectively. For cryptic species such as some species of reptiles (Bates et al. 2014; McGrath et al. 2015), there is often a distinct lack of high-quality records of these animals' occurrences within their natural environments (Böhm et al. 2013; Tolley et al. 2016), even within areas specifically designated for conservation (Ferreira et al. 2011; Venter et al. 2008; Zielinski 2001).

In South Africa, the Kruger National Park (KNP) is home to approximately 126 reptile species (Bates et al. 2014; Branch 1998; Pienaar 1978). The presence of reptiles promotes ecological diversity within KNP, and more broadly southern Africa, as many reptile species are likely to have important ecological roles or carry out ecologically beneficial functions within a variety of habitats and ecosystems (Trimble & Aarde 2014). Overall, reptiles comprise approximately 14% of vertebrate species within KNP (Parr, Woinarski & Pienaar 2009), and the conservation of these animals is essential for maintaining diversity within this important protected area (Gascon et al. 2015; Parr et al. 2009; Venter et al. 2008).

The KNP biological reference collection houses thousands of preserved specimens across a wide variety of taxa and includes hundreds of individual reptiles collected within the park over the past 80 years. The collection also includes an extensive electronic database that catalogues each specimen along with its respective biological and locality information where available. This collection places KNP among the best sampled protected areas in South Africa (and probably in Africa) for reptiles (Bates et al. 2014). However, the very nature of such reference collections is that sampling intensity and objectives vary over time, with earlier sampling efforts focused primarily on compiling inventory lists and collecting reference material. As such, the KNP biological reference collection database for reptiles was never intended as a systematic survey across all habitats and reflecting all patterns of occurrence within the park. In recent years, however, the need for spatially explicit species occurrence

data sets to inform modern conservation tools requires that the data from biological reference collections should be co-opted into conservation analyses. Accordingly, there is a need to critically evaluate such existing data sets to understand any inherent patterns of bias they possess.

In this study, we (1) collate and synthesise available occurrence data for reptile species in KNP from reference collections, museum databases and literature sources; (2) assess patterns of geographic and taxonomic biases within this data set; and (3) evaluate whether areas of spatial bias are environmentally representative of KNP as a whole, including under-sampled regions.

## Methods

### Reptile occurrence data

We collated reptile locality and occurrence data from literature sources, museum and reference collection databases, a virtual museum platform, citizen science sightings from social media platforms and field data gathered under various teaching, monitoring and inventorying exercises by the Organization for Tropical Studies (Table 1). Additionally, two of the authors (J.M.B. and B.M.) provided 151 novel records from personal observations in KNP (listed as 'this study'). In total, we collated 14 533 records, but after georeferencing these to match locality descriptions and removing duplicates across sources, we had a final data set of 7118 records representing 126 reptile species occurring in KNP. This data set is available upon request from SANParks Scientific Services.

### Coverage biases

We summarised reptile species occurrence data to identify geographic and taxonomic biases in coverage across KNP. By carrying out regression analyses, we tested if reptile families were evenly represented across KNP by comparing the relationship between the number of occurrences for each reptile family to the extent of the geographical areas (in km<sup>2</sup>) surrounding those occurrences of each reptile family (i.e. the area of the minimum convex polygon enclosing all

**TABLE 1:** Sources of reptile occurrence records used in this study.

Source	Number of records	Purpose
Museums	1440	
Ditsong National Museum of Natural History	499	Specimen collection
Skukuza Biological Reference Collection	941	Inventory listing, specimen collection, distribution mapping
Literature	1908	
Pienaar (1978)	1523	Distribution mapping
Jacobsen (1989)	331	Distribution mapping
Other literature	54	Research
Organization for Tropical Studies	134	Teaching and learning exercises
Virtual museums	3430	
ReptileMAP	3430	Distribution mapping
Sightings	47	Distribution mapping
This study	159	Distribution mapping
$\Sigma$	<b>7118</b>	-

occurrences). Next, we assessed if residual values of each reptile family fell within the range of mean  $\pm$  standard deviation of all residual values to identify which families were spatially over- or under-represented (i.e. above or below the range) across KNP in our sample data set.

To evaluate the proportion of KNP for which reptile occurrence data exist and quantify the extent of unsampled areas, we divided KNP into equal-sized grid cells at 1 km  $\times$  1 km, 2 km  $\times$  2 km, 4 km  $\times$  4 km and 9 km  $\times$  9 km (pentad scale) resolutions, respectively. These resolutions allowed us to identify patterns of geographic sampling bias across a range of biologically appropriate spatial resolutions. However, the 1 km  $\times$  1 km resolution was preferred for most analyses. This resolution subjectively offered the best trade-off between the spatial error associated with historical records of occurrence data (Newbold 2010) and the relatively small spatial scale at which many reptiles utilise landscapes (Fischer, Lindenmayer & Cowling 2004; Price, Kutt & McAlpine 2010). We plotted reptile occurrences across the grid cells of each spatial resolution by using Quantum Geographic Information System (QGIS) version 3.4 (QGIS Development Team 2018) and counted the number of occurrences per grid cell. By carrying out regression analyses, we also tested whether a relationship exists between the numbers of reptile occurrences recorded within each grid cell and the proximity of those grid cells to the nearest publicly accessible infrastructure areas of KNP (defined here as all camps, gates, picnic sites and public roads) at the finest spatial resolution (1 km  $\times$  1 km).

### Are sampled areas representative of the Kruger National Park as a whole?

We downloaded environmental and infrastructural data layers at a spatial resolution of 1 km  $\times$  1 km to represent the overall environmental space of KNP. These included 20 bioclimatic layers representing current climate (1970–2000) and elevation from the Worldclim database (<http://www.worldclim.org>), soil type classifications of South Africa from the International Soil Resource and Information Centre (ISRIC; <https://www.isric.org>), vegetation type classifications of South Africa from the South African National Biodiversity Institute (SANBI; [www.bgis.sanbi.org](http://www.bgis.sanbi.org)) and infrastructural layers for publicly accessible camps, gates, picnic sites and roads within KNP from South African National Parks (SANParks; <http://dataknpsanparks.org/sanparks>). We also generated 'slope', 'aspect' and 'distance to water bodies' layers for KNP by using ArcGIS version 10.4 (ESRI 2016), resulting in a total of 27 representative layers for the environmental space of KNP.

To reduce the effects of spatial autocorrelation between layers, we performed a principal component analysis by using R version 3.5.3 (R Core Team 2018) to summarise the layers into 27 new, uncorrelated principal component layers. We retained the first six principal component layers as representatives of the overall environmental variability of KNP as they cumulatively represented 85% of all modelled variation, which we selected as an effective stopping point as per

Jackson (1993). We tested if sampled areas within KNP (i.e. grid cells containing at least one reptile occurrence) were environmentally representative of KNP as a whole. To do this we separated each of our six principal component layers into new layers representing (1) sampled areas and (2) unsampled areas and compared the kernel density estimates of each pair per component via six separate two-sample Kolmogorov–Smirnov tests.

### Ethical considerations

This article followed all ethical standards for a research without direct contact with human or animal subjects.

## Results

### Summary of occurrence data

Our database contained 7118 reptile occurrence records, unevenly distributed across 60 lizard species, 59 snake species, 6 testudine species and 1 crocodylian species (Table 2). As such, the majority of occurrences were of squamates (lizards: 48% of all records; snakes: 41% of all records), with the less speciose testudine and crocodylian groups having less representation (8% and 3% of all records, respectively). This was not the case at the species level where the Nile crocodile (210 records) and the leopard tortoise (232 records) ranked only below the rainbow rock skink (242 records) for species with the highest numbers of occurrence records in our data set. The number of occurrence records per reptile family was positively related to the number of species per said family (Linear regression analysis:  $F_{1,17} = 28.45$ ,  $p < 0.01$ ,  $R^2 = 0.63$ ). The uneven distributions of records across reptile families were likely

**TABLE 2:** Summary of records of reptile occurrences within the Kruger National Park.

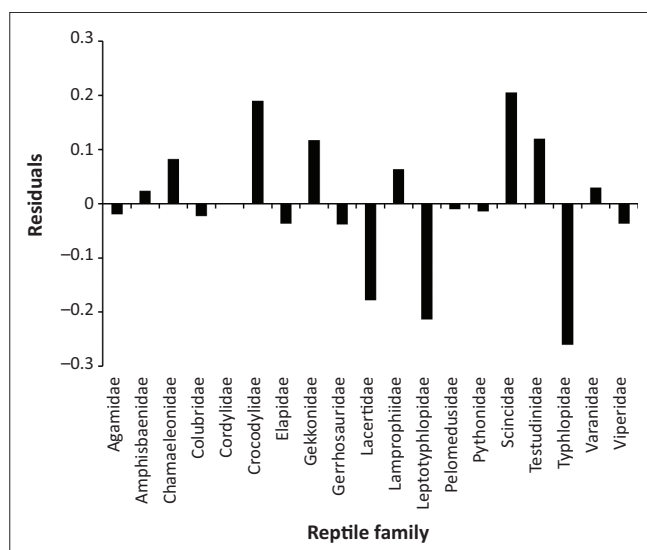
Group	Number of species	Number of records	Percentage of total records
<b>Lizards</b>	<b>60</b>	<b>3434</b>	<b>48</b>
Agamidae	3	214	3
Amphisbaenidae	7	195	3
Chamaeleonidae	1	154	2
Cordylidae	9	206	3
Gekkonidae	14	708	10
Gerrhosauridae	4	344	5
Lacertidae	6	272	4
Scincidae	14	1096	15
Varanidae	2	245	3
<b>Snakes</b>	<b>59</b>	<b>2944</b>	<b>41</b>
Colubridae	10	621	9
Elapidae	7	403	6
Lamprophiidae	29	1233	17
Leptotyphlopidae	5	221	3
Pythonidae	1	125	2
Typhlopidae	3	130	2
Viperidae	4	211	3
<b>Chelonians</b>	<b>6</b>	<b>530</b>	<b>8</b>
Pelomedusidae	3	230	3
Testudinidae	3	300	4
<b>Crocodylians</b>	<b>1</b>	<b>210</b>	<b>3</b>
Crocodylidae	1	210	3
<b>Σ</b>	<b>126</b>	<b>7118</b>	<b>100</b>

present as a product of collection bias and the specific combination of species occurring within KNP rather than being solely because of collection bias on its own.

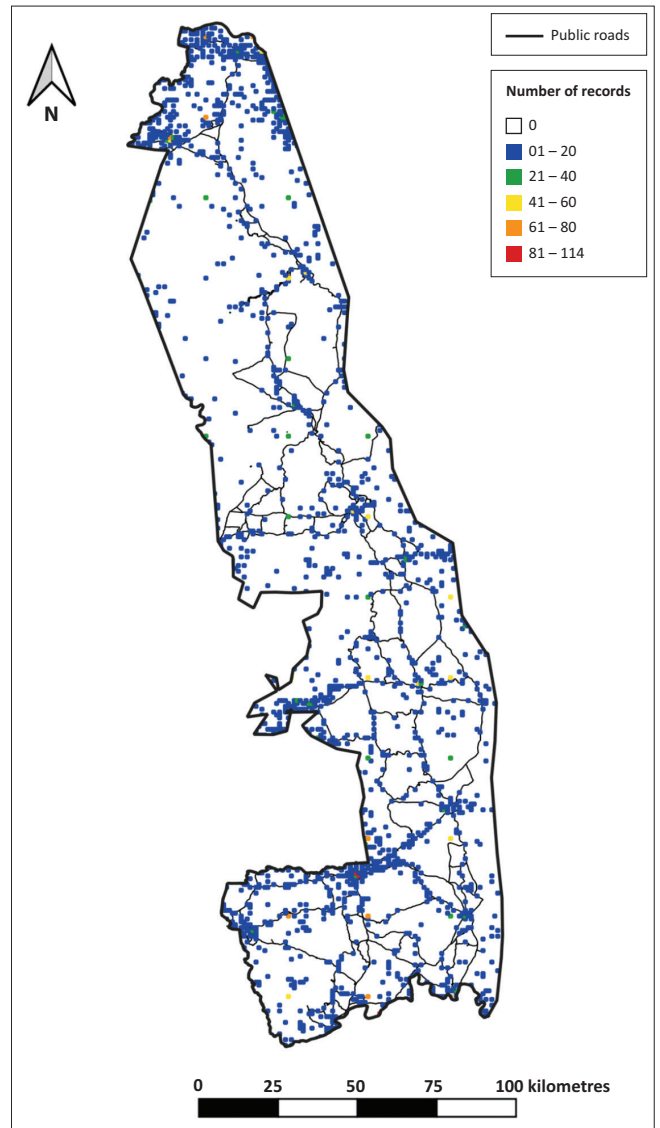
### Coverage biases

Representation based on spatial coverage was unevenly distributed across reptile families in KNP. We found a significant positive relationship between the cumulative number of records and the cumulative extents of the areas encompassing records of each reptile species per reptile family (linear regression analysis:  $F_{1,17} = 76.60$ ,  $p < 0.01$ ,  $R^2 = 0.81$ ). We identified reptile families that appeared to be significantly under-represented (Lacertidae, Leptotyphlopidae and Typhlopidae) and those that were over-represented (Crocodylidae and Scincidae) geographically across KNP (Figure 1), providing evidence of taxonomic sampling bias at the family level within our data set.

At the biologically appropriate spatial resolution of 1 km × 1 km, we found that only 1751 of 21 761 grid cells (8%) contained any reptile occurrence records at all (Figure 2). Moreover, 52% of these grid cells contained only a single record (911 grid cells; Figure 3). We found that as the numbers of records per grid cell increased, the numbers of grid cells containing records decreased (regression analysis:  $F_{1,114} = 9.34$ ,  $p < 0.01$ ,  $R^2 = 0.08$ ). This pattern held true at resolutions of 2 km × 2 km and 4 km × 4 km, respectively, but was not present at 9 km × 9 km (Figure 3), demonstrating that geographic sampling bias is strongest at fine resolutions but weakens as resolution becomes coarser. We also found a significant relationship between the number of records present within each grid cell and its proximity to publicly accessible human infrastructure within the park (regression analyses: public roads  $-F_{1,1749} = 6.75$ ,  $p < 0.01$ ,  $R^2 = 0.06$ ; camp sites and picnic spots  $-F_{1,1749} = 9.01$ ,  $p < 0.01$ ,  $R^2 = 0.10$ ; Figure 4). As the distance to infrastructure increased, the frequency of recorded reptile occurrences per



**FIGURE 1:** Linear regression residuals demonstrating coverage biases across reptile families within the Kruger National Park. Families with positive residuals occur more frequently than expected based on family-level species richness, although families with negative residuals are considered under-sampled.



**FIGURE 2:** Spread of reptile occurrence records across the Kruger National Park at a 1 km × 1 km spatial resolution.

grid cell significantly decreased, providing evidence of sampling bias towards publicly accessible areas.

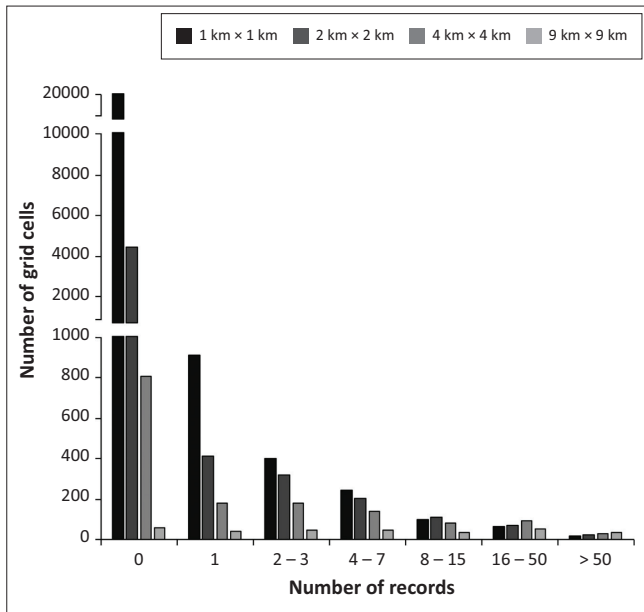
### Environmental representation of sampled areas

At the spatial resolution of 1 km × 1 km, sampled areas of KNP were not representative of the full range of environmental space of KNP as a whole (Figure 5). The results of six separate Kolmogorov–Smirnov tests showed that there were significant differences in environmental variability between sampled and unsampled areas across each of the six principal components representing the overall environmental space of KNP ( $D = 0.06–0.27$ ,  $p < 0.01$  in all cases). Grid cells containing records of reptile occurrences were thus not statistically representative of the overall ecological variability of KNP.

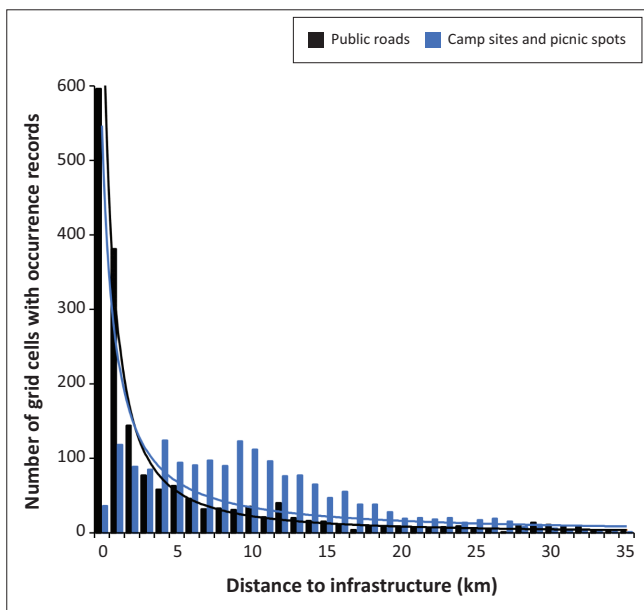
### Discussion

At a fine spatial resolution ecologically relevant to reptiles, occurrence data for reptile species in KNP are geographically





**FIGURE 3:** Frequency distributions of the numbers of Kruger National Park reptile occurrence records per grid cell at 1 km × 1 km, 2 km × 2 km, 4 km × 4 km and 9 km × 9 km resolutions.



**FIGURE 4:** Distances of grid cells (1 km × 1 km) containing records to nearest publicly accessible infrastructure (camp sites, picnic spots and public roads) within the Kruger National Park.

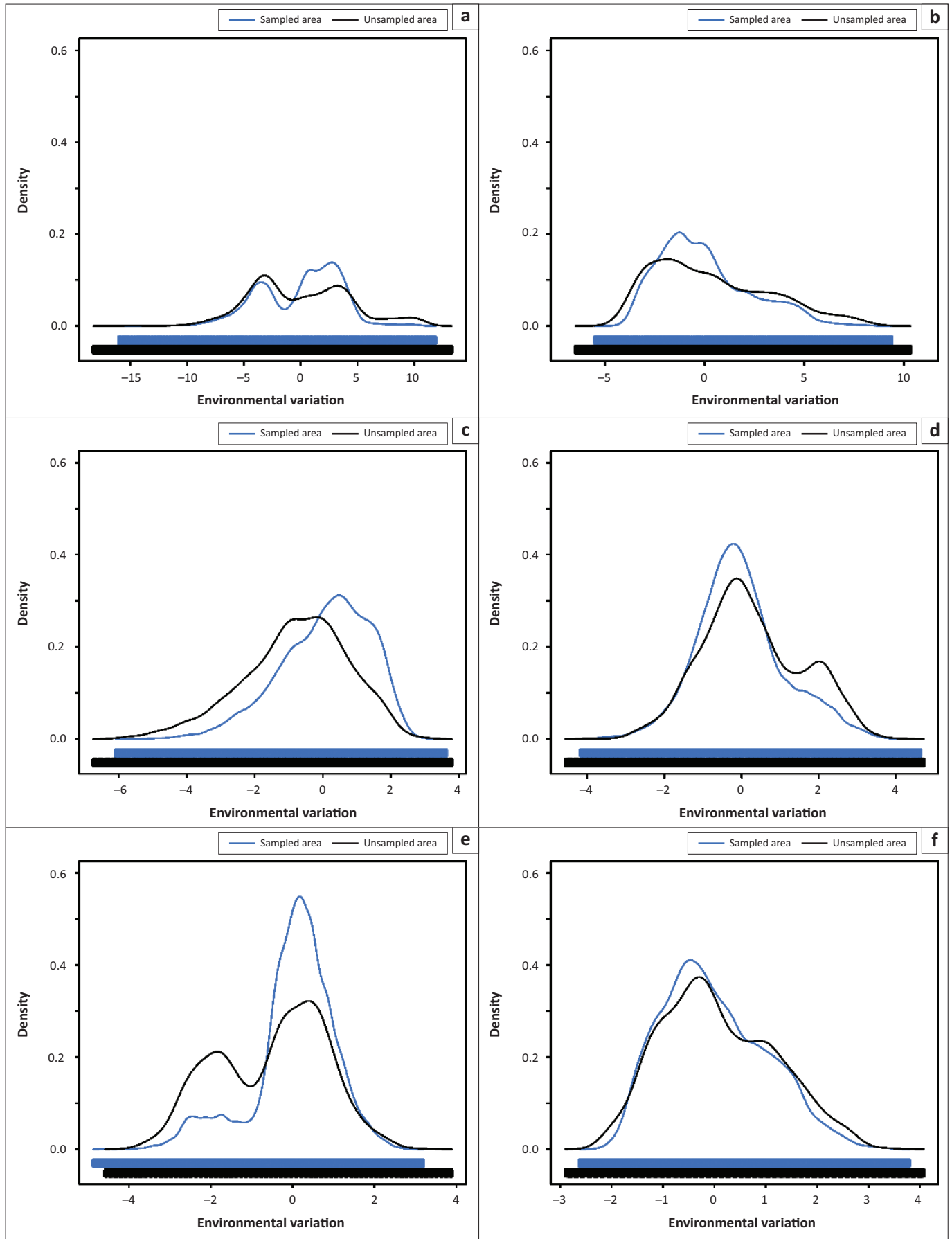
and taxonomically biased. As a consequence of overall data deficiency, representation within our reptile occurrences data set varied, with highly detectable reptile species and families having had significantly more records than those with comparatively lower detectability. Moreover, the majority of reptile occurrence data were associated with human infrastructure. Approximately 68% of all records occurred in close proximity (< 2 km) to publicly accessible human infrastructure areas in KNP. Unsurprisingly, grid cells associated with major tourist camps and surrounding areas were considerably better sampled than the remainder of the park. Importantly, sampled areas were not representative of the complete range of environmental variability across KNP. This

suggests that regions of the park that comprised unique environmental space are not represented in the current data set.

Spatial sampling biases associated with human infrastructure are common in biological sampling data sets (Newbold 2010). Most notably, presence-only data sets derived from atlas projects, citizen science data and museum records are typically susceptible to geographic bias in collection effort (Botts et al. 2011; Geldmann et al. 2016; Reddy & Davalos 2003; Zielinski 2001;). Geographic bias in collection effort is often present within these data sets as a result of sampling being inhibited in certain areas but facilitated in others (Botts et al. 2011; Pyke & Ehrlich 2010). For example, some areas may be difficult to sample because of extreme weather conditions, rough terrain, the presence of dangerous animals, distance from roads or restricted access (Bird et al. 2014; Freitag et al. 1998). Conversely, other areas facilitate more complete sampling by providing ease of access and associated increased visitation. In this context, sampling intensity within certain areas is likely to be dramatically lower in comparison with that of less restrictive areas that offer greater accessibility. This is certainly the case in KNP where publicly accessible infrastructural areas have increased human presence and accessibility from staff and visitors alike in comparison with the remainder of the park. Consequently, our data set seemingly represents areas of high sampling intensity rather than true biological patterns.

Areas of high sampling intensity seldom represent the full range of environmental space and ecological factors associated with determining species distributions (Tolley et al. 2016). Several studies have found substantial differences in climate between well- and under-sampled areas (see Botts et al. 2011; Kadmon et al. 2004; Martinez & Wool 2006; Reddy & Davalos 2003; Stockwell & Peters 1999), with many of these highlighting the significance of climatic biases towards assessments of the true biological distributions of species. Botts et al. (2011) found that sparse sampling effort in areas away from human infrastructure resulted in incomplete representations of amphibian distributions across South Africa. Here, where we have encountered similar geographical sampling biases with our reptile data set to those encountered by Botts et al. (2011), we similarly conclude that our data set is unlikely to reflect the complete range of real-world distributions of reptile species across KNP.

The biased nature of our data set has important implications for SANPark's management of reptiles in KNP. Despite being perhaps the most comprehensive collation of reptile occurrence data for KNP to date, the use of this data set for informing robust conservation management decisions would need to be considered with caution. Because of the large variation in geographical sampling intensity within our data set and the associated biases within the underlying occurrence data, it would be inappropriate to use this data set in its current form within the context of spatial planning for species conservation management. Because a large proportion of our data was not collected explicitly for the purposes of mapping species distributions, the biological patterns as presented



**FIGURE 5:** Overlap in environmental variation between sampled and unsampled grid cells within the Kruger National Park at the 1 km × 1 km resolution for (a) principal component 1, (b) principal component 2, (c) principal component 3, (d) principal component 4, (e) principal component 5, and (f) principal component 6. Bars beneath kernel density plots represent the range of environmental variation between sampled and unsampled grid cells.

within our data set may confound comparisons among species, or comparisons of a particular species' abundance in different habitat types, across environmental gradients, or across time series (Bird et al. 2014; Fischer et al. 2004). However, should the geographic biases be minimised or reversed without concurrently increasing other forms of bias (Botts et al. 2011), this data set could become an important resource for KNP conservation management.

Minimising the biases within our data set could be achieved by targeted sampling of data-deficient areas in KNP. Although the majority of 1 km × 1 km grid cells in KNP could benefit from supplemental sampling, priority should be given to grid cells that contain no data and are distant from publicly accessible areas that are steadily subject to human visitation. In particular, we recommend that the mopane woodlands-dominated north-western region of the park (i.e. areas demarked as ecozones P and P1 as per SANParks 2016) should be targeted for additional sampling. The majority of this region is poorly sampled, with most of its grid cells containing no reptile records. Moreover, this region has few public roads and is largely lacking in human visitation. Here, we recommend that an approach emulating the Karoo Biogaps project led by SANBI (Main et al. 2019) should be implemented, in which specific grid cells are selected as sites in which to extensively collect occurrence records on the basis of a statistical sampling design. This method of grid cell selection would involve the use of a statistical algorithm (such as Latin hypercube sampling) that seeks to maximise coverage across the region although minimising the total number of grid cells to be sampled. Traversal to grid cells targeted for sampling could be facilitated through the use of management roads unavailable to the general public. Together with public tourist roads, this offers the widest range of vehicle access across KNP. Although these roads do not cover the full extent of KNP, their usage can alleviate some of the challenges associated with inaccessibility for many grid cells and offers a feasible option towards future sampling campaigns.

Systematically sampling for reptile occurrences in targeted and supplementary areas within KNP will substantially improve upon the overall coverage and comprehensiveness of available data. It is important to note however that the challenges associated with reptile sampling, such as low detectability (McGrath et al. 2015), may result in underestimations of species richness and occurrences at specific sites. False absences as a result of underestimations can falsely inform on species' performance within monitoring frameworks, including those relating to thresholds of potential concern, and may result in incorrect assignments of conservation priority (Botts et al. 2011; Ferreira et al. 2011). Compiling complete inventories for targeted grid cells will thus be vitally important, but this may require several sampling trips to ensure that comprehensive species lists are compiled. Such an approach would be unavoidably costly, time-consuming and could delay investigations of the statuses of reptile species within the park.

In the meantime, other options are available to fill gaps in sampling within KNP reptile data. Over the last two decades, an increasing number of studies have used species distribution models (SDMs) to extrapolate spatially explicit predictions of the distributional ranges of species (Bird et al. 2014; Stockwell & Peters 1999). Species distribution models predict environmental suitability for species, which can be used to infer species' presence or absence within a given area (Guisan et al. 2013; Hurlbert & Jetz 2007; Kery 2011). These types of models are typically referred to as ecological niche models as the predictions produced are based on statistical relationships between species occurrences and environmental descriptor variables (Guisan et al. 2013; Kadmon et al. 2004). Importantly, studies have found that SDMs based on biased data with limited occurrences can produce strong models with accurate predictions (e.g. Pearson et al. 2007; Syfert et al. 2013), if the underlying biases are corrected for during model production and high-quality predictor variables are available.

Studies that aim to identify sources of data bias, particularly within presence-only data sets, offer invaluable insights into bias correction within the context of SDMs (Syfert et al. 2013). By understanding the sources of bias, it may be possible to correct historical and current population distributions modelled through SDM frameworks using mathematically inferred or experimentally determined bias correction factors. A good example of this is the use of visibility bias correction factors on aerial census data to improve the accuracy of geographical distributions and population size estimates of large herbivores in KNP (see Redfern et al. 2002). Potential SDM frameworks for KNP reptile species should seek to correct for the proximity of reptile occurrences to publicly accessible areas within the park. Overcoming the challenges associated with this bias may require innovative solutions; however, if implemented correctly, such a framework could offer a feasible approach towards obtaining meaningful reptile distribution information for use within conservation management in KNP.

## Conclusion

We sought to collate occurrence records for KNP reptile species and provide a quantitative assessment of the sampling biases within these data. We have shown that at biologically relevant resolutions, KNP is largely data deficient for reptile occurrences, with existing data being geographically biased towards publicly accessible areas. We further show that sampled areas were not environmentally representative of KNP as a whole and from this, we conclude that our data set does not provide a true reflection of real-world reptile species distributions across KNP. Because the majority of the data within our database were not explicitly collected with species mapping in mind, additional sampling is needed to reverse the biases present. We recommend that future sampling efforts should target historically poorly sampled regions in the park that are distant from publicly accessible locations. Finally, we suggest that in the meantime SDMs may offer a more feasible approach for use within conservation management decision-making relating to reptile species within KNP.

## Acknowledgements

The authors thank the following institutions for providing reptile occurrence data: the Animal Demography Unit (University of Cape Town), the Ditsong National Museum of Natural History, Organization for Tropical Studies (OTS) and SANParks. We acknowledge and thank all of the various collectors for their valuable contributions towards expanding the the KNP reference collection over the last 80 years. They further thank OTS, including Bernard Coetzee and Laurence Kruger for providing logistical support and assistance in the field. They also thank Robin Maritz for input towards the discussion.

## Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

## Authors' contributions

B.M. conceptualised this study and obtained data provided by D.W.P., D.R.C.T. and G.Z. J.M.B. collated the data and performed all analyses. J.M.B. and B.M. co-wrote the manuscript, with additional assistance from D.R.C.T.

## Funding information

The authors acknowledge the Ada and Bertie Levenstein Foundation, Merseta and the National Research Foundation (NRF grant no. 99186) for providing funding that facilitated the completion of this study.

## Data availability statement

Reptile occurrence data used in this study are available from South Africa National Parks.

## Disclaimer

All the views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any affiliated agency of the authors.

## References

- Bates, M.F., Branch, W.R., Bauer, A.M., Burger, M., Marias, J., Alexander, G.J. et al., 2014, *Suricata 1: Atlas and red list of the reptiles of South Africa, Lesotho, and Swaziland*, South Africa National Biodiversity Institute, Pretoria.
- Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J. et al., 2014, 'Statistical solutions for error and bias in global citizen science datasets', *Biological Conservation* 173(1), 144–154. <https://doi.org/10.1016/j.biocon.2013.07.037>
- Böhm, M., Collen, B., Baillie, J.E., Bowles, P., Chanson, J., Cox, N. et al., 2013, 'The conservation status of the world's reptiles', *Biological Conservation* 157(1), 372–385. <https://doi.org/10.1016/j.biocon.2012.07.015>
- Botts, E.A., Erasmus, B.F.N. & Alexander, G.J., 2011, 'Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning', *Biodiversity Conservation* 20(1), 119–139. <https://doi.org/10.1007/s10531-010-9950-6>
- Branch, W.R., 1998, *Field guide to snakes and other reptiles of southern Africa*, Struik, Cape Town.
- ESRI, 2016, *ArcGIS Desktop 10.4*, Environmental Systems Research Institute, Redlands, CA, viewed 28 February 2018, from <https://www.arcgis.com/index.html>.

- Ferreira, S., Deacon, A., Sithole, H., Bezuidenhout, H., Daemane, M. & Herbst, M., 2011, 'From numbers to ecosystems and biodiversity: A mechanistic approach to monitoring', *Koedoe* 53(2), 1–12. <https://doi.org/10.4102/koedoe.v53i2.998>
- Fischer, J., Lindenmayer, D.B. & Cowling, A., 2004, 'The challenge of managing multiple species at multiple scales: Reptiles in an Australia grazing landscape', *Journal of Applied Ecology* 41(1), 32–44. <https://doi.org/10.1111/j.1365-2664.2004.00869.x>
- Franklin, J., 2010, *Mapping species distributions: Spatial inference and prediction*, Cambridge University Press, Cambridge.
- Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S., 1998, 'Testing for potential survey bias: The effect of roads, urban areas and nature reserves on a southern African mammal data set', *Animal Conservation Forum* 1(2), 119–127. <https://doi.org/10.1111/j.1469-1795.1998.tb00019.x>
- Gascon, C., Brooks, T.M., Contreras-MacBeath, T., Heard, N., Konstant, W., Lamoreux, J., et al., 2015, 'The importance and benefits of species', *Current Biology* 25(10), R431–R438. <https://doi.org/10.1016/j.cub.2015.03.041>
- Geldmann, J., Heilmann-Clausen, J., Holm, T.E., Levinsky, I., Markussen, B., Olsen, K. et al., 2016, 'What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements', *Diversity and Distributions* 22(11), 1139–1149. <https://doi.org/10.1111/ddi.12477>
- Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T. et al., 2013, 'Predicting species distributions for conservation decisions', *Ecology Letters* 16(12), 1424–1435. <https://doi.org/10.1111/ele.12189>
- Hurlbert, A.H. & Jetz, W., 2007, 'Species richness, hotspots, and the scale dependence of range maps in ecology and conservation', *Proceedings of the National Academy of Sciences* 104(33): 13384–13389. <https://doi.org/10.1073/pnas.0704469104>
- Jackson, D.A., 1993, 'Stopping rules in principal components analysis: A comparison of heuristic and statistical approaches', *Ecology* 74(8), 2204–2214. <https://doi.org/10.2307/1939574>
- Jacobsen, N.G.H., 1989, 'A herpetological survey of the Transvaal', PhD thesis, Dept. of Biological Sciences, University of Natal.
- Kadmon, R., Farber, O. & Danin, A., 2004, 'Effects of roadside bias on the accuracy of predictive maps produced by bioclimatic models', *Ecological Applications* 14(2), 40–413. <https://doi.org/10.1890/02-5364>
- Kery, M., 2011, 'Towards the modelling of true species distributions', *Journal of Biogeography* 38(4), 617–618. <https://doi.org/10.1111/j.1365-2699.2011.02487.x>
- Main, D., Tensen, L., Gihring, K., Bronner, G., Aboul-Hassan, N., Blanckenberg, M. et al., 2019, 'Unravelling the taxonomy and distribution of two problematic small mammal genera in the Karoo biome', *African Zoology* 54(3), 125–135. <https://doi.org/10.1080/15627020.2019.1628661>
- Martinez, J.L. & Wool, D., 2006, 'Sampling bias in roadsides: The case of galling aphids on *Pistacia* trees', *Biodiversity & Conservation* 15(7), 2109–2121. <https://doi.org/10.1007/s10531-004-6685-2>
- McGrath, T., Guillera-Aroita, G., Lahoz-Monfort, J.J., Osborne, W., Hunter, D. & Sarre, S.D., 2015, 'Accounting for detectability when surveying for rare or declining reptiles: Turning rocks to find the grassland earless dragon in Australia', *Biological Conservation* 182(1), 53–62. <https://doi.org/10.1016/j.biocon.2014.11.028>
- Newbold, T., 2010, 'Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models', *Progress in Physical Geography* 34(1), 3–22. <https://doi.org/10.1177/0309133309355630>
- Parr, C.L., Woinarski, J.C.Z. & Pienaar, D.J., 2009, 'Cornerstones of biodiversity conservation? Comparing the management effectiveness of Kruger and Kakadu National Parks, two key savanna reserves', *Biodiversity Conservation* 18(1), 3643–3662. <https://doi.org/10.1007/s10531-009-9669-4>
- Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T., 2007, 'Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar', *Journal of Biogeography* 34(1), 102–117. <https://doi.org/10.1111/j.1365-2699.2006.01594.x>
- Pienaar, U.D.V., 1978, *The reptile fauna of the Kruger National Park*, National Parks Board of South Africa, Pretoria.
- Price, B., Kutt, A.S. & McAlpine, C.A., 2010, 'The importance of fine-scale savanna heterogeneity for reptiles and small mammals', *Biological Conservation* 143(1), 2504–2513. <https://doi.org/10.1016/j.biocon.2010.06.017>
- Pyke, G.H. & Ehrlich, P.R., 2010, 'Biological collections and ecological/environmental research: A review, some observations and a look to the future', *Biological Reviews* 85(2), 247–266. <https://doi.org/10.1111/j.1469-185X.2009.00098.x>
- QGIS Development Team, 2018, *QGIS Geographic Information System version 3.4. Open Source Geospatial Foundation Project*, viewed 28 February 2018, from <https://www.qgis.org/en/site/index.html>.
- R Core Team, 2018, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, viewed 05 March 2018, from <https://www.r-project.org>.
- Reddy, S. & Davalos, L., 2003, 'Geographic sampling bias and its implications for conservation priorities in Africa', *Journal of Biogeography* 30(11), 1719–1727. <https://doi.org/10.1046/j.1365-2699.2003.00946.x>
- Redfern, J.V., Viljoen, P.C., Kruger, J.M. & Getz, W.M., 2002, 'Biases in estimating population size from an aerial census: A case study in the Kruger National Park, South Africa: Starfield Festschrift', *South African Journal of Science* 98(9), 455–461.
- SANParks, 2016, *Kruger National Park official guide*, Jacana Media, Melville.



- Stockwell, D. & Peters, D., 1999, 'The GARP modelling system: Problems and solutions to automated spatial prediction', *International Journal of Geographical Information Science* 13(2), 143–158. <https://doi.org/10.1080/136588199241391>
- Syfert, M.M., Smith, M.J. & Coomes, D.A., 2013, 'The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models', *PLoS One* 8(2), e55158. <https://doi.org/10.1371/journal.pone.0055158>
- Tolley, K.A., Alexander, G.J., Branch, W.R., Bowles, P. & Maritz, B., 2016, 'Conservation status and threats for African reptiles', *Biological Conservation* 204(Part A), 63–71. <https://doi.org/10.1016/j.biocon.2016.04.006>
- Trimble, M.J. & Aarde, R.J., 2014, 'Amphibian and reptile communities and functional groups over a land use gradient in a coastal tropical forest landscape of high richness and endemism', *Animal Conservation* 17(5), 441–53. <https://doi.org/10.1111/acv.12111>
- Venter, F.J., Naiman, R.J., Biggs, H.C. & Pienaar, D.J., 2008, 'The evolution of conservation management philosophy: Science, environmental change and social adjustments in Kruger National Park', *Ecosystems* 11(2), 173–192. <https://doi.org/10.1007/s10021-007-9116-x>
- Zielinski, P., 2001, 'Dealing with uneven recording effort in regional atlas projects on the distribution of amphibians and reptiles', *Acta Zoologica Cracoviensia* 44(2), 85–92.