Ecography

ECOG-03944
Monsarrat, S., Boshoff, A. and Kerley, G. 2018. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. – Ecography doi: 10.1111/ecog.03944

Supplementary material

Appendix 1: Constraints faced by European travellers in South Africa

An understanding of the constraints faced by observers in the early historical period is key to understanding how their movements, and hence their observations, are likely to be spatially biased. Most expeditions were undertaken using wagons, typically pulled by large teams of oxen (Joubert, 1995). Travellers also hired staff and carried livestock, provisions and merchandise for barter. For example, in his second trip into the interior parts of southern Africa, the French explorer and naturalist François Le Vaillant described a company of 19 people, 3 wagons, 52 oxen, 13 dogs, 11 goats, 3 horses, 3 cows and numerous barter articles (Le Vaillant, 1796). The trips were difficult owing to the lack of any form of roads, of which construction only started towards the end of the 19th century, and the rough nature of the terrain (Mentzel, 1921). To reach distant inland settlements, wagons had to make their way across country, over rivers, and through passes to cross the mountain ranges following previous wagon tracks or cutting across open habitat (Joubert, 1995) (e.g. Johan August Wahlberg on 10th October 1841: "For a good part of the way, we travelled over the veld, on no road at all" (Craig and Hummel, 1992)). Travellers had no fear of starvation, since they had opportunities for hunting game, but water was a major concern (Mentzel, 1921). Even though travellers were able to store some amount of water in leather gourds and jars, they would ideally try to get a daily access to freshwater, as illustrated by the numerous mentions in their journals of the efforts in finding overnight stopovers close to rivers (e.g. Craig and Hummel, 1992). Should travellers fail to find freshwater for two or three days, they and their livestock were liable to die of thirst (e.g. "During the whole day we had nothing but a dry and burning desert to traverse. After dinner, two of my oxen, exhausted by thirst and fatigue, dropped down, and I was under the necessity of leaving them behind" (Le Vaillant, 1796, p.213)). Thus, as Skead (2011:x) puts it, "not all areas were visited or settled by early observers. For example, some places were avoided by them because they were 'off the beaten track' or perhaps because there was no surface water there to supply their needs and those of their horses and livestock". A consequence of these environmental constraints is the spatial and environmental bias that we observe in the historical written records of species occurrence collected by these early observers.

Appendix 2: Test of Complete Spatial Randomness

We tested if the occurrence records are spatially biased using simulation envelopes, a well-established technique for testing hypotheses about spatial point patterns (Baddeley et al., 2014). Simulation envelopes are based on computing a summary function of the point pattern, such as Ripley's K function (Dixon, 2002), to test complete spatial randomness (CSR), i.e. whether the observed events are consistent with a homogeneous Poisson process. Ripley's K function was computed from the point data and plotted as the solid black curve in Fig. A2.1. Then 39 simulated, random point patterns were generated according to CSR, the K functions of each simulated pattern were computed, and the maximum and minimum K values were plotted as the limits of the gray-shaded envelope in Fig. A2.1. If the pattern process arises from CSR, then the observed K(t) function should be similar to the K(t) from the simulated data.

The solid black line in Fig. A2.1 lies above the expected value of K(t) (dotted red line) for all distances (r) between 1 and 2 degrees, indicating that the occurrence records (K_{obs}) significantly deviate from a random process (K_{theo}) and that the occurrence records are clustered compared to what would be expected from a homogeneous Poisson process.

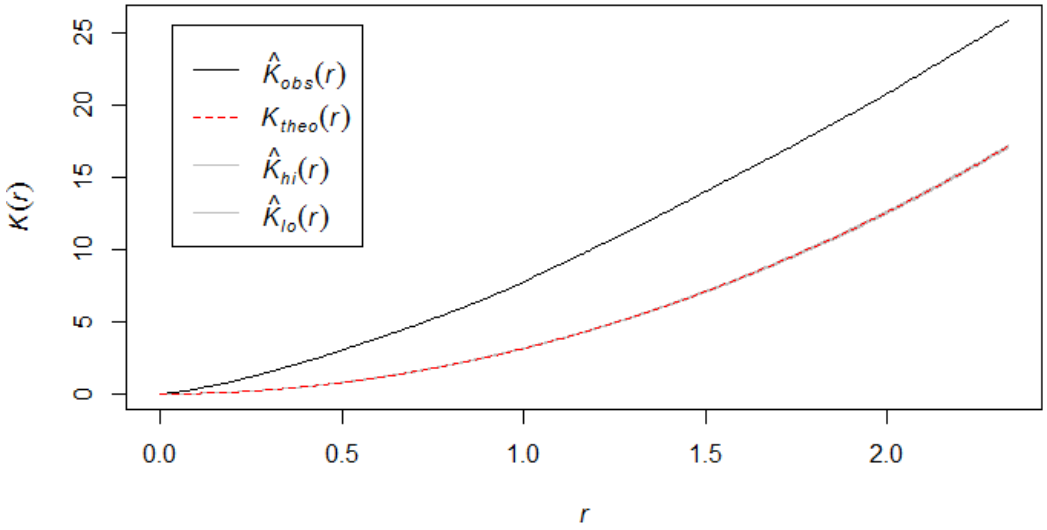

Figure A2.1. Analysis of spatial point pattern using simulation envelopes. Solid lines represent Ripley's K function computed from the data pattern from Fig. 1. Shading indicates the envelope of values obtained from 39 simulations of complete spatial randomness (CSR), K_{hi} nd K_{lo} representing the confidence limits around the theoretical expectation (note that the shading is so close to K_{theo} that it is barely visible on the figure). The dashed red line shows the theoretical value for CSR.

Appendix 3: Temporal biases in historical written records and predictive ability of the model for different time periods

We plotted the spatial distribution of historical written records for 6 periods between 1720 and 1920 to identify temporal biases in the records (Fig. A3.1). As the distribution of settlements changed over time, so did the accessibility of the landscape. We calculated and plotted the accessibility index AI for each time period, taking into account the establishment date of European settlements (Table A3.1) (i.e. only the settlements established before a given date were used to calculate AI in the time-period that ends at this date). We evaluated the model's ability to predict observer's presence for each period using the method described in the method section in the main text (Fig. A3.2, Table A3.2).

Table A3.1 Establishment year of European settlements in the study area from the 17th to the 20th centuries, based on Floyd (1960).

1652 Cape Town	1820 Bathurst	1844 Victoria West	1856 Aberdeen	1852 Jagersfontein	1875 Zastron
1679 Stellenbosch	1820 Grahamstown	1846 Burghersdorp	1856 Montagu	1854 Kroonstad	1876 Vrede
1680 Simonstown	1822 Fort Beaufort	1846 Bloemfontein	1857 Outsdhoorn	1856 Boshof	1877 Senekal
1745 Malmesbury	1823 Philipolis	1849 Fauresmith	1858 Adendorp	1859 Bethlehem	1878 Frankfort
1746 Swellendam	1825 Durbanville	1849 Harrismith	1858 Cathcart	1859 Jacobsdal	1878 Vredefort
1786 Graaf-Reinet	1825 Somerset East	1849 Smithfield	1861 McGregor	1861 Reddersburg	1880 Dewetsdorp
1795 Tulbagh	1830 Colesberg	1887 Kuruman	1871 Kokstad	1862 Edenburg	1882 Parys
1804 Uitenhage	1835 King William'sTown	1851 Calvinia	1871 Knysna	1863 Bethulie	1889 Bothaville
1806 Port Elizabeth	1835 Winburg	1852 Mahikeng	1872 East London	1863 Rouxville	1890 Reitz
1806 Clan William	1838 Riversdale	1853 Queenstown	1873 Barkly East	1868 Ladybrand	1891 Ficksburg
1811 Caledon	1838 Napier	1853 Robertson	1875 Sterkstroom	1872 Heilbron	1891 Petrusburg
1811 George	1840 Wellington	1853 Carnarvon	1876 Ngcobo	1873 Bultfontein	1891 Villiers
1812 Griquatown	1840 Piketberg	1854 Bedford	1880 Vryburg	1873 Hoopstad	1892 Fouriesburg
1816 Cradock	1841 Villiersdorp	1854 Ceres	1882 Morreesburg	1875 Brandfort	1892 Koffiefontein
1818 Beaufort West	1842 Prince Albert	1854 Jansenville	1864 Bloemhof	1875 Lindley	1892 Thaba'Nchu
1819 Worcester	1843 Richmond	1856 Stanford	1870 Christiana	1875 Wepener	

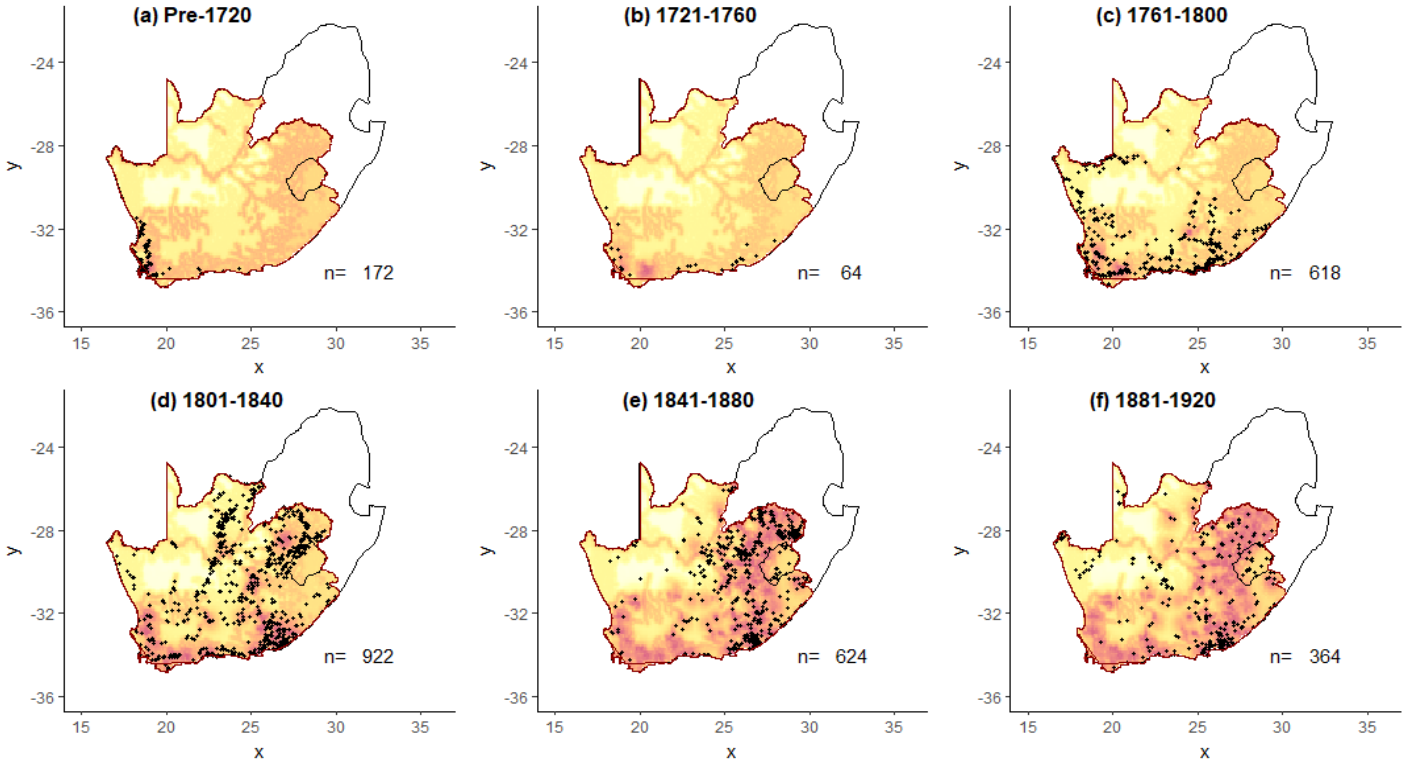

Figure A3.1 Distribution of historical written records in the study area and Accessibility Index (AI) (a) Pre-1720 (b) 1721-1760 (c) 1761-1800 (d) 1801-1840 (e) 1841-1880 (f) 1881-1920. Shades of red indicate progressively higher accessibility as predicted by the model. The study area is indicated with a red outline. The number of records (n) for each period is indicated on the bottom right of each panel.

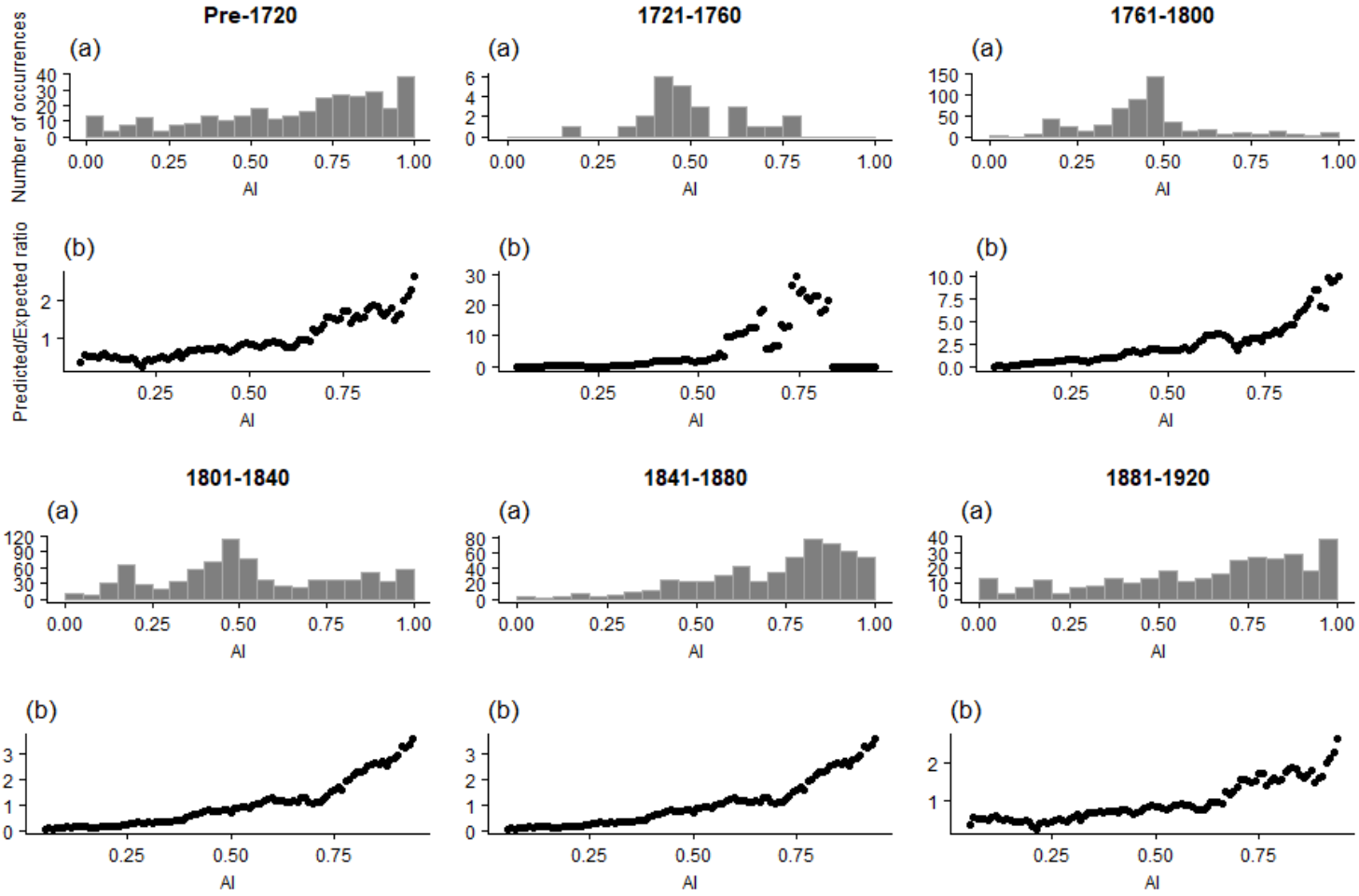

Figure A3.2. Model performance plots for the six periods of time presented in Fig. A3.1. For each plot, (a) is the histogram of number of occurrences against the predicted Accessibility Index (AI); (b) is the Predicted/Expected (P/E) curve (see method section in the main text for more details).

Table A3.2. Value of continuous Boyce index (B_{cont}) and Pearson's correlation coefficient (ρ) for each period of time considered in Fig. A3.1.

Period	B_{cont}	ρ
Pre-1720	0.939	0.39
1721-1760	0.973	0.008
1761-1800	0.980	-0.141
1801-1840	0.951	0.2
1841-1880	0.990	0.912
1881-1920	0.952	0.828

The high values of continuous Boyce index (B_{cont}) indicate that the model's predictions differ from random distribution of the observed presences across the prediction gradient for all time periods. However, for the periods between 1721 and 1840, the model's ability to predict observer's presence is poor, as evidenced by low values of Pearson's correlation coefficient (ρ).

These results can be interpreted in the light of South Africa colonial history:

Pre-1720. Cape Town was the first colonial settlement established by the Dutch in 1652. Until the late 1670s, the European settlements did not extend beyond the plains around the Cape peninsula, principally because of conflicts with the neighboring Khoikhoi (Giliomee and Mbenga, 2007). In 1720, only three towns (Cape Town, Simonstown and Stellenbosch) had been established, all located around the Cape Peninsula. The distribution of records (Fig. A3.1a) and the relatively good predictive ability of AI in the pre-1720 period is in accordance with the behavior of early settlers, who concentrated around the Cape, with few exploration journeys into the interior of the country.

1721-1800. The late 1710s saw a rapid European expansion over a large territory when the Dutch East India Company allowed farmers to develop beyond the borders of the colony. Initially known as the trekboers, these farmers dispersed away from the Cape in search of better pastures. The area of European occupation grew almost tenfold between 1703 and 1780, and by 1800, the territory occupied by Europeans covered 286,000 km^2 in the south and south-western parts of South Africa (Giliomee and Mbenga, 2007).

The observed shift in the distribution of records towards the north and east of the study area is consistent with this history of European colonisation. However, the model performs poorly at predicting the distribution of these records (Table A3.2). With only seven settlements established before 1800 in the study area (Table A3.1), the Settlement Proximity Index (SPI) component of AI appears to be a poor predictor of the distribution of occurrence records at that period. The histograms of number of occurrences vs AI (graphs (a) in Fig. A3.2) show that the number of occurrences increases up to AI=0.5 and decreases afterwards. The 0.5 value of AI (which is calculated as the mean of SPI and WPI) must correspond to locations where the value of Water Proximity Index (WPI) is at its maximum but SPI is close to zero. During this period of exploration, it thus seems that freshwater availability alone is a better predictor of occurrence distribution than AI.

1801-1840. The British took over in 1806 and in the 1830s, many Afrikaners, called the Voortrekkers, left the then Cape Colony en masse to establish new settlements in what became the Transvaal and the Orange Free State republics, this in response to political issues. The establishment of European farms and settlements in these regions paved the way for more literate travellers to penetrate this part of the country, providing us with the first written descriptions of the interior of South Africa.

The model performs poorly at predicting the distribution of presences in that period. This can be explained by the fact that some settlements were already important for European travellers before their official date of establishment. For example, Thaba Nchu was the last settlement to be established in the 19th century (Table A3.1). However, a station of the Wesleyan Missionary Society was established in Thaba Nchu in 1833 (Watson, 1977) and Voortrekkers and other travellers had been congregating there since the 1830s. This illustrates the lag in the building of European settlements and their official date of establishment. This also justifies why AI does not quite fit the distribution of presences in the study area before the 1840s, until lots of these settlements were made official.

1841-1920. By this time, following years of frontier wars with native communities, the English and Voortrekkers occupied most of the study area. The observed decrease in the number of records after 1840, with notably fewer records in the western part of the country, is unlikely to be caused by a decrease in the number of observers at that period in South Africa (if anything, the number of available sources should be higher in the more recent past). This pattern is most likely explained by the collapse of populations of large mammals from over hunting and loss of habitat that occurred in the 19th century (Boshoff and Kerley 2015). The model performs well in predicting the distribution of records, which suggests that both SPI and WPI are relevant components of AI for that period.

Overall, this analysis brings a more detailed picture of the temporal dynamic of observer's distribution and the relevance of AI for predicting sampling bias at different time periods. The results suggest that, in the exploration phase of South Africa, when settlements were already in place but not officially established, freshwater alone is a better predictor of traveler's distribution. In the most recent past, adding settlement proximity in the model increases the ability of AI to predict occurrence records, and when the entire period is considered, AI is a good predictor of sampling bias (see results in the main text). This suggests that, for a given time period, locations that are known to be repeatedly visited but will only be made official later should be included in the model, in order to draw a better picture of the accessibility in the area at that time. In general, when considering any period of time, care must be given to select the appropriate features that then constrain the movement of observers.

Appendix 4: Discussion on including terrain or barriers in the accessibility model

The accessibility map could have been built using a different method such as path-cost analysis, accounting for physical barriers that affect the permeability of the landscape to the movement of human observers. However, elements related to the very peculiar behavior of early travelers motivated us to not include terrain or barriers in the map of accessibility.

Early travelers had strong motivations to reach a particular place, whether for religious missions, exploration, naturalist, military or industrial interests, etc. They would manage to travel through difficult terrain and overcome major barriers (e.g. rivers, mountains) on the way, as some accounts testify:

- Francois Le Vaillant, in his second voyage to South Africa, describes how he perilously crossed a large river in order to search for elephants on the opposite bank: *"I shall here give a particular account of the celebrated instance of fool-hardiness, which was nothing less than to cross with my fire-arms, baggage and attendants, a considerable river swelled by inundations, in order to encamp on the opposite side"* (Le Vaillant, 1796:239).
- J.A. Wahlberg in his voyage to South Africa and Namibia/Botswana 1838-1856: "*On a steep hill called Hale Huug the oxen were unable to haul the wagon up until we had offloaded over half of the contents and laboriously carried them up"* (Craig and Hummel, 1992:33).
- W.J. Burchell in 1822: "*… having on our right some high mountains in the distance and before us an exceedingly large table mountain […] From the distance and spot at which it was viewed, it appeared inaccessible, being surrounded on all sides by a precipice; but experience teaches that however steep and lofty a mountain may appear, its summit should not be pronounced inaccessible until its ascent have been attempted on every side*" (Burchell, 1822:84).

Adding terrain in the model (with the effect of reducing the Accessibility Index AI where the slope is high) was actually found to decrease the ability of AI to predict observer's presences (difference in continuous Boyce index ΔB_{cont}= - 0.236, and difference in Pearson's correlation coefficient $\Delta\rho$= - 0.11). Some records were collected in grid cells where the slope is very high, highlighting the fact that observers would use those areas regardless of the terrain they encountered there. Overall, the decision of observers to follow steep routes or to cross dangerous rivers might have resulted in a longer travel time and greater perils. However, the motivation to explore new, poorly accessible, areas could have been strong enough to overcome these difficulties. Ultimately, every point in the environment was "accessible" to travelers with the adequate motivation, if it was not dramatically far from basic resources like settlements and water.

In addition, the resolution of our study, with ca. 10 km wide grid cells, does not allow us to take into account very localized barriers. Even faced with steep slopes, travelers had the ability to follow sinuous passes that are not easy to identify at a low resolution. Trying to include localized barriers such as cliffs or mountain peaks in our model at that large spatial scale would likely provide irrelevant outcomes.

Appendix 5: Sensitivity analysis

The comparison of spatial bias in the model and the historical occurrence data is dependent on several parameters (when building the model and testing its performance), whose values were chosen based on our knowledge of observer's behaviour and environmental constraints. However, some of these choices are inferred rather than known, which raises questions of the relevance of the chosen parameter values. It is thus important to test the sensitivity of the approach to different parameters values, in order to validate its robustness. Additionally, it is also important to investigate the relative importance of WPI and SPI in predicting sampling density. To investigate this, we looked at the effect of setting either SPI or WPI to 0 on the model performance, and we calculated changes in the predictive performance of the model when using different values of parameters. We modified each parameter separately, keeping all other parameters equal as in the original model (model 1), and compared values of the continuous Boyce index (*Bcont*), a measure of the ability of the model to predict observer's presences (Hirzel et al., 2006) (see main text for details on this method), and Pearson's correlation coefficient (ρ), which provides information on the linear correlation between the number of observed occurrences and predicted AI, to obtain Δ*Bcont* and Δρ (Table A2.1).

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

From the original model above (see main text for the description of model parameters), we tested the following changes in parameters:

1) Setting either the settlement proximity index (SPI) or the water proximity index (WPI) to 0 (partial models)

2) Different sizes of kernel widths h_s and h_w in the Gaussian kernel function used to calculate the proximity indexes *SPI* and *WPI.*

3) Triangular kernel instead of a Gaussian kernel function to calculate the proximity indexes *SPI* and *WPI* (model 2).

$$AI(x_i) = \frac{1}{2}\left(\left(\underbrace{1 - \frac{dist_s}{max(dist_s)}}_{SPI}\right) + \left(\underbrace{1 - \frac{dist_w}{max(dist_w)}}_{WPI}\right)\right) \quad (2)$$

Table A5.1. Value of continuous Boyce index (B_{cont}) and Pearson's correlation coefficient (ρ) for different values of parameters used to build the accessibility model. ΔB_{cont} and Δρ are the differences between the B_{cont} and ρ of the original model and those of the models built with different parameter values.

	Value of parameters	B_{cont}	ρ	Δ B_{cont}	Δρ
Full model	Hs=40km, Hw=10km, Density function=Gaussian kernel, Pseudo-absences buffer size=20km	0.995	0.93	-	-
Partial models	WPI (SPI set to 0)	0.959	0.75	-0.036	-0.18
	SPI (WPI set to 0)	0.977	0.63	-0.018	-0.30
Kernel widths (km)	Hs=10	0.989	-0.14	-0.006	-1.07
	Hs=80	0.990	0.86	-0.005	-0.07
	Hw=5	0.996	0.75	0.001	-0.18
	Hw=20	0.985	0.88	-0.010	-0.05
Shape of density function	Triangular kernel SPI	0.944	0.86	-0.051	-0.07
	Triangular kernel WPI	0.983	0.07	-0.012	-0.86
	Triangular kernel SPI & WPI	0.902	0.21	-0.093	-0.72

Model performance when setting either the WPI or SPI component of AI to 0 indicates that the full model consistently performs better in predicting sampling density (Table A2.1; Fig. A2.1). Notably, WPI and SPI models only have a weak linear correlation between the frequency of observed occurrences and predicted AI (measured by the correlation coefficient ρ, Table A4.1) and slightly lower values of B_{cont}.

We find that ΔBcont in Table A2.1 is consistently very low (<0.1), suggesting that the model discrimination ability is robust to changes in parameters values. However, the correlation coefficient shows a marked decrease for low values of h_s and h_w, indicating that the scale at which the feature influences the movement of observers is an important parameter. Setting it too low increases the risk of type II errors, i.e. the model will predict low accessibility in areas that are in fact accessible. Also, the Gaussian function seems to be a better descriptor of the influence of features on the movement of observers than the triangular kernel.

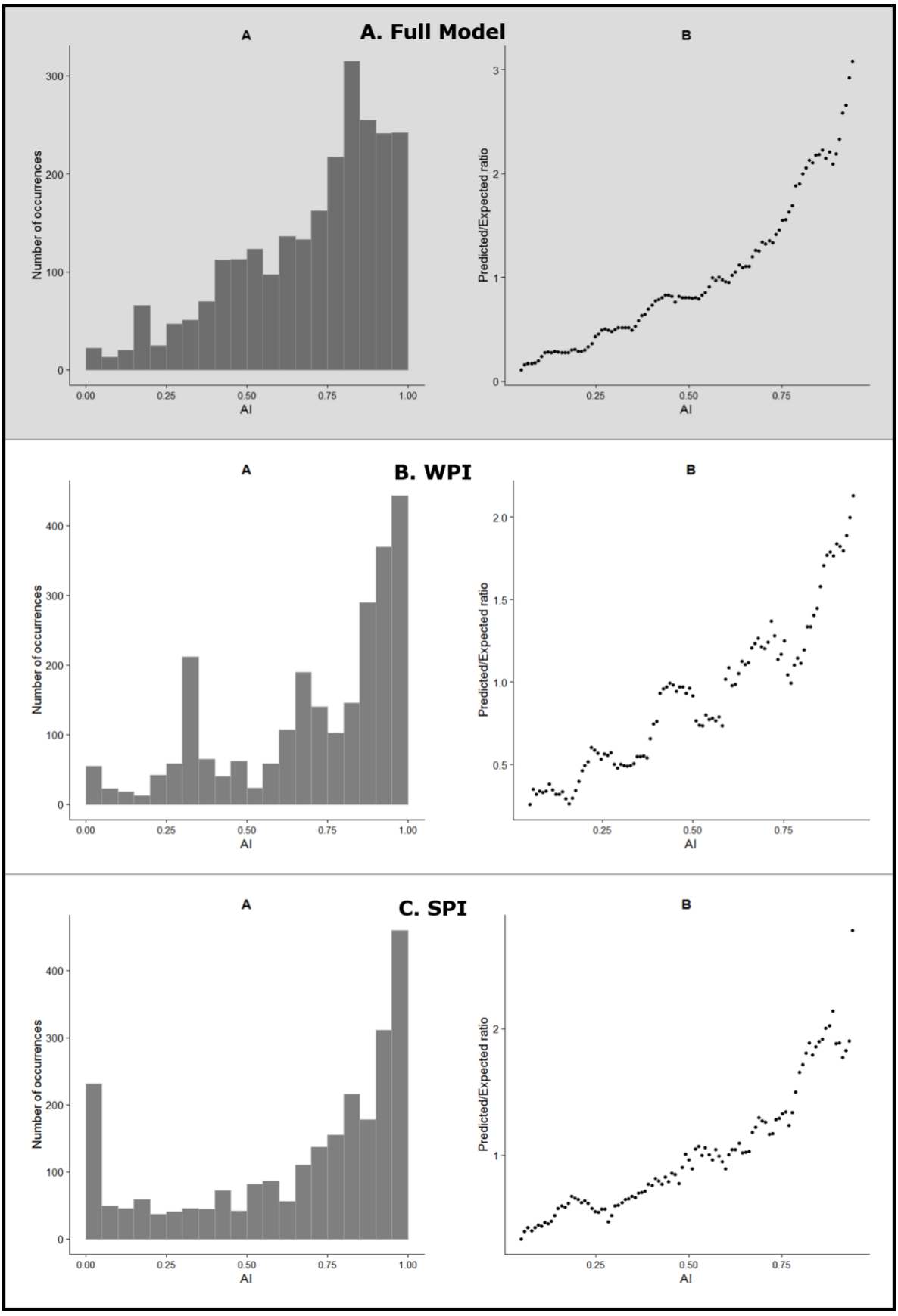

Figure A5.1 Model performance plots for (A) the full model, (B) the WPI model (SPI set to 0) and (C) the SPI model (WPI set to 0). For each plot, (a) is the histogram of number of occurrences against the predicted Accessibility Index (AI); (b) is the Predicted/Expected (P/E) curve. Each point is calculated as the ratio of frequency of evaluation points predicted by the model (P) and the expected frequency from a random distribution across the study area (E) for the corresponding AI class. A straight curve indicates an ideal model with perfect predictive ability.

Appendix 6: Model evaluation based on the density of records

In addition to testing the model's ability to predict observer's *presence* (main text), we tested its ability to predict the *density* of records. We first plotted the number of records per cell vs the accessibility index (AI) and calculated Spearman's correlation coefficient between the two variables. We then built a zero-truncated Poisson regression model and mapped the residuals to identify possible spatial patterns. The plot of number of records vs AI indicates a poor fit between the two variables (Fig. A6.1, Spearman's correlation coefficient=0.21), notably because of the existence of many cells with 5 or fewer records in high AI areas (overall, more than 95% of cells have 5 or less records). The map of residuals does not show a strong spatial pattern (Fig. A6.2). The higher residuals appear to be in areas where the number of records is excessively high (e.g. near towns such as Cape Town, Port Elizabeth, i.e. harbors where the time of residence would have been high hence the number of recorded mammals is particularly high too), where the model is not able to predict such density of records.

Overall, the accessibility index is a better predictor of the observer's presence (Boyce index=0.995, Pearson's correlation coefficient=0.93, see Fig. 4 in the main text) than of the density of records per cell (this appendix). In the context of providing a tool to address sampling biases in spatial analyses, and in species distribution models in particular, we believe that being able to predict observer's presence is a more important achievement than predicting the density of records. However, this comparison provides useful information for future work aiming at assessing survey quality and the reliability of historical biodiversity data.

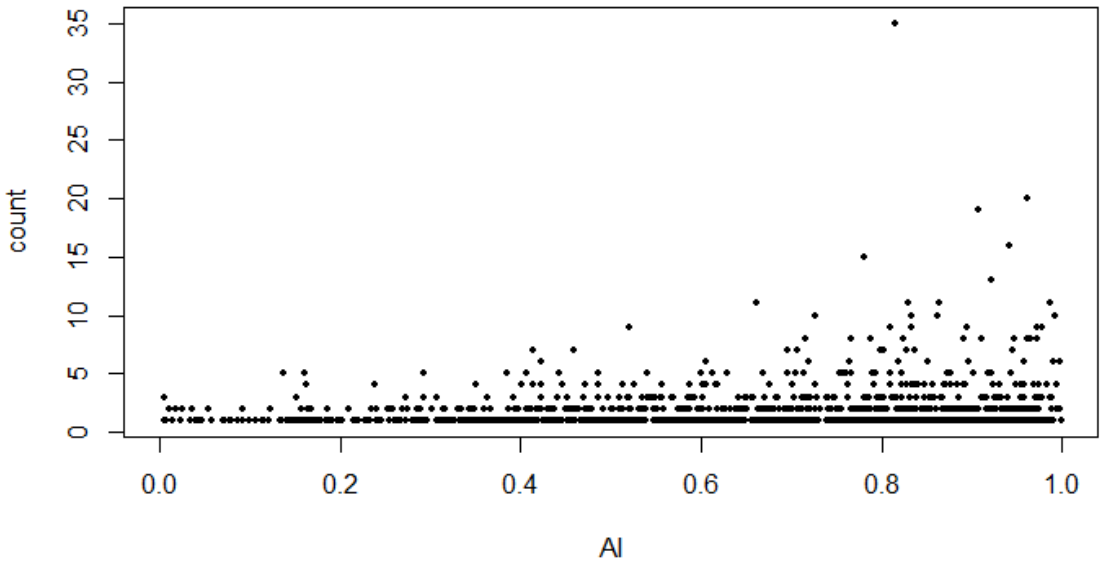

Figure A6.1 Plot of number of records per cell vs the Accessibility Index (AI). Spearman's correlation coefficient between these two variables is 0.21.

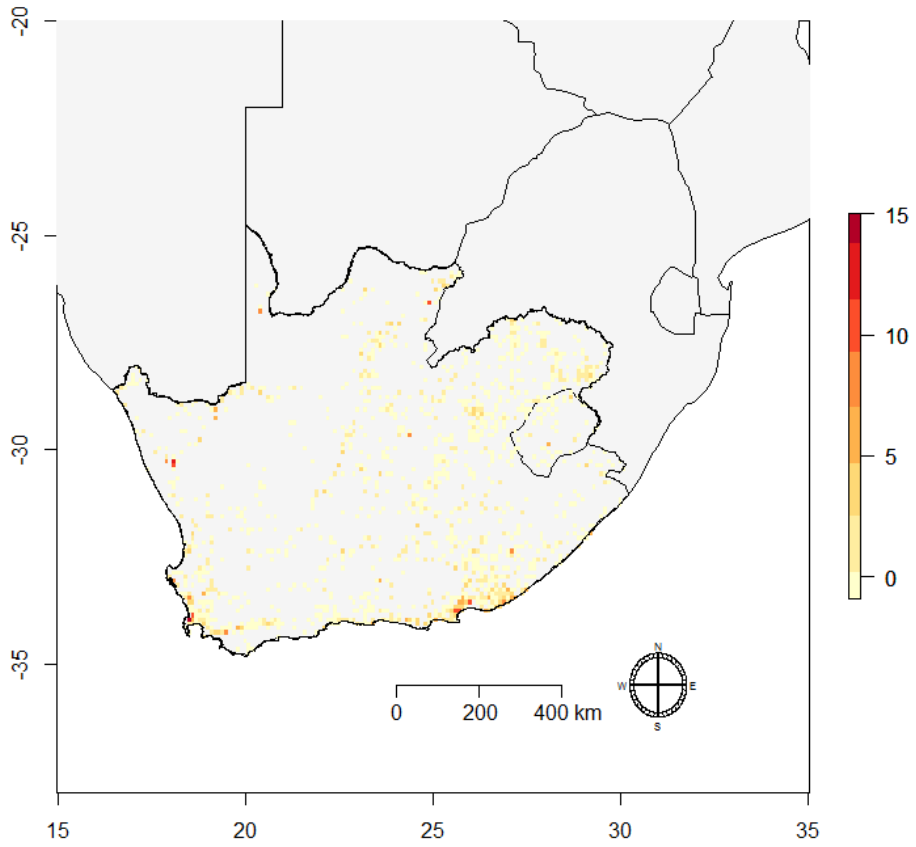

Figure A6.2 Map of residuals from a zero-truncated Poisson regression model between the number of records per cell and the accessibility index (AI). Darker values indicate larger residuals, i.e. where the model performed poorly at predicting the observed density of records.

REFERENCES

Baddeley, A., Diggle, P.J., Hardegen, A., Lawrence, T., Milne, R.K., Nair, G., 2014. On tests of spatial pattern based on simulation envelopes. Ecol. Monogr. 84, 477–489.

Burchell, W.J., 1822. Travels in the Interior of Southern Africa, Travels in the Interior of Southern Africa. Longman, Hurst, Rees, Orme, and Brown.

Craig, A., Hummel, C., 1992. Johan August Wahlberg: Travel Journals (and Some Letters) South Africa and Namibia/Botswana, 1838-1856. Van Riebeeck Society.

Dixon, P.M., 2002. Ripley's K function. Encycl. Environmetrics 3, 1796–1803.

Floyd, T.B., 1960. Town Planning in South Africa, accessed from https://fr.scribd.com/document/291256918/Chronological-Order-of-Town-Establishment-in-South-Africa. ed. Shuter & Shooter, Pietermaritzburg.

Giliomee, H.B., Mbenga, B., 2007. New History of South Africa. Tafelberg.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199, 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Joubert, B., 1995. An historical perspective on animal power use in South Africa, in: Animal Power in South Africa: Empowering Rural Communities. Starkey, P (ed), Gauteng: Development Bank of Southern Africa, pp. 125–138.

Levaillant, F., 1796. New travels into the interior parts of Africa : by the way of the Cape of Good Hope, in the years 1783, 84 and 85. London :Printed for G.G. and J. Robinson.

Mentzel, O.F., 1921. A geographical and topographical description of the Cape of Good Hope - With translation of original t.p.: A complete and authentic geographical and topographical description of the famous and (all things considered) remarkable African Cape of Good Hope / by O.F. Mentzel. Glogau : C.F. Günther, 1785-87. Van Riebeeck Society, Cape Town.

Skead, C.J., 2011. Historical Incidence of the Larger Land Mammals in the Broader Western and Northern Cape. (eds Boshoff, A., Kerley, G.I.H. & Lloyd, P). Centre for Conservation Ecology, Nelson Mandela Metropolitan University.

Watson, R.L., 1977. Missionary Influence at Thaba Nchu, 1833-1854: A Reassessment. Int. J. Afr. Hist. Stud. 10, 394. https://doi.org/10.2307/216734