Appendix S1: Supplementary Material for "Automated surface feature selection using SALSA2D: An illustration using Elephant Mortality data in Etosha National Park."

Ecosphere

L.A.S Scott-Hayward[a], M.L. Mackenzie[a], C.G. Walker[b], G. Shatumbu[c], W. Kilian[c] and P. du Preez[d]

[a]School of Mathematics and Statistics, University of St Andrews, KY16 9LZ, Fife, Scotland; [b]Department of Engineering Science, University of Auckland, 70 Symonds Street, Auckland, New Zealand; [c]Etosha Ecological Institute, PO Box 6, Okaukuejo via Outjo, Ministry of Environment, Forestry and Tourism, Namibia; [d]African Wildlife Conservation Trust, PO box 97401, Windhoek, Namibia

[a]ORCiD: LSH (0000-0003-3402-533X), MLM (0000-0002-8505-6585)

ARTICLE HISTORY

Compiled August 16, 2023

1. Finding the largest residual

- Find nearest candidate knot location (of the legal knots remaining and ignoring the already selected knots) to each data point (both presence and pseudo-absence locations). Note that "nearest" is calculated based on whichever distance metric the model uses. Figure S1 shows the neighbourhood around each knot.
- Sum the observed counts within each knot region
- Make predictions to the pseudo absence grid and sum the estimated intensity

CONTACT L.A.S. Scott-Hayward. Email: lass@st-andrews.ac.uk

within each knot region

- Calculate the absolute residual ($|(O - E)|$)
- Find the 10 knot regions with the largest score. These become the candidates for the exchange/move step.

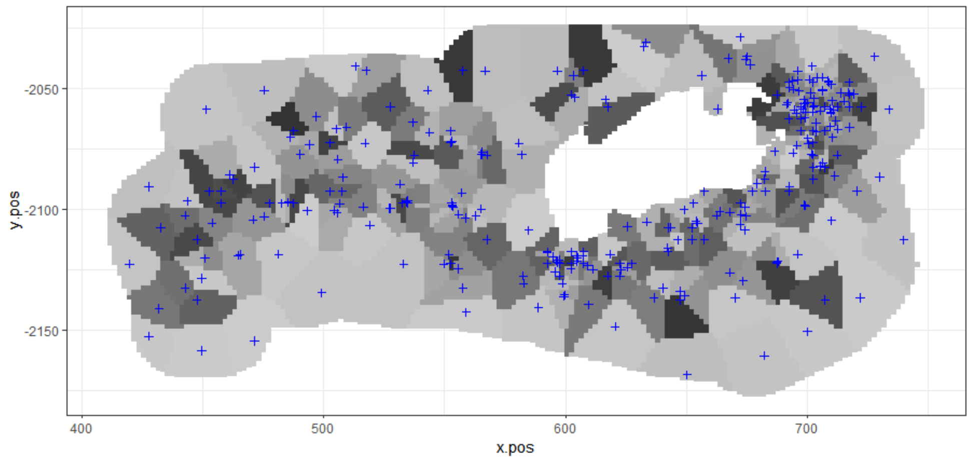

Figure S1. Figure showing the knot locations (blue crosses) and the colour shows the nearest knot on the pseudo absence grid.

2. Pseudo-absence Selection

- Grid spacings trialled: 5, 4, 3, 2, 1.5, 1.25 and 1 km
- SALSA2D specification
  - knot grid: all non-duplicated presence locations
  - start knot number: 10, 20, 30, 40
  - min knots and max knots equal to start knot number.
  - distance metric: Euclidean
  - basis: Gaussian and exponential
- Fit models for each specification of grid, start knots and basis
- Evaluate the log-likelihood
- Select the coarsest resolution after which, an increase in resolution makes little difference to the likelihood.

Figure S2 shows the log-likelihood scores for the different parameterisations. The vertical dashed line indicates the best grid resolution; 2km^2.

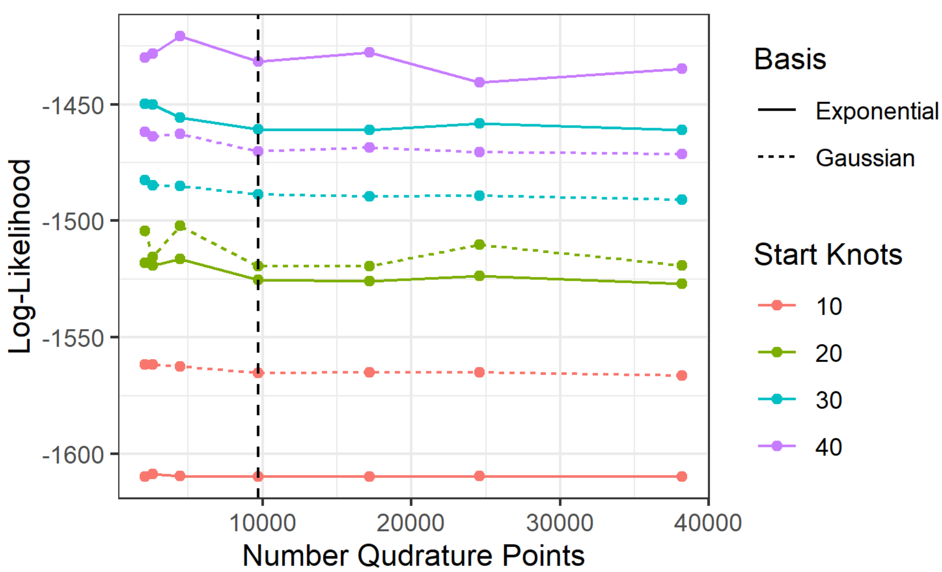

Figure S2. Figure showing the convergence of the log-likelihood for for different spatial resolutions across multiple SALSA2D parameterisations.

3. Estimated Intensities

The following figures show the best models from the two different methodological frameworks and the four different parameterisations. The best models are selected using BIC for SALSA2D and AIC$_c$ weights for model averaging. Since, the two frameworks are compared using the log-likelihood, Figure S3 shows the best log-likelihood selected models.

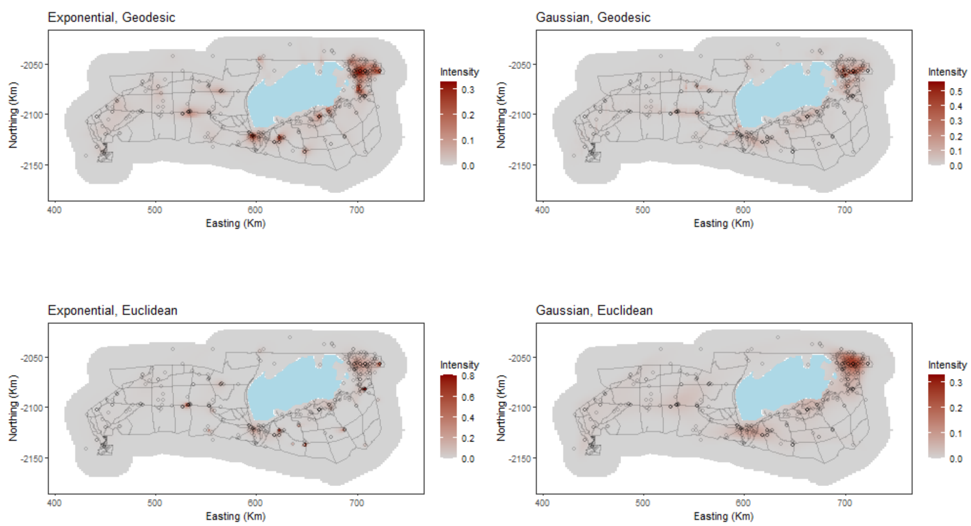

Figure S3. Figure showing fitted intensity surfaces for the four SALSA2D models selected using log-likelihood.

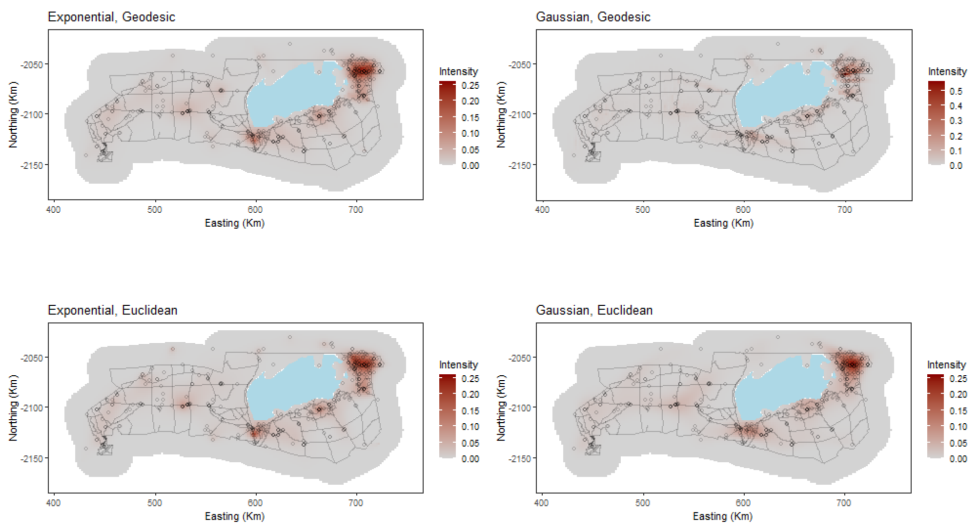

Figure S4. Figure showing the best model averaged outputs from the four different parametrisations and selected using AIC$_c$ model weights.

SALSA2D outputs

Distance Type	Basis	Start Knots	End Knots	LogLik	BIC	Time (min)
Euclidean	Exponential	5	8	-1528.5	3130.6	2.3
Euclidean	Exponential	10	8	-1509.1	3091.9	2.1
Euclidean	Exponential	15	8	-1517.0	3107.7	2.7
Euclidean	Exponential	20	10	-1504.6	3101.4	4.6
Euclidean	Exponential	25	12	-1486.3	3083.1	21.1
Euclidean	Exponential	30	19	-1445.0	3065.0	22.9
Euclidean	Exponential	35	22	-1438.7	3080.0	21.6
Euclidean	Exponential	40	27	-1407.7	3064.1	46.7
Euclidean	Exponential	45	28	-1376.8	3011.5	49.9
Euclidean	Exponential	50	32	-1374.5	3043.8	66.7
Euclidean	Exponential	55	40	-1318.0	3004.4	73.3
Euclidean	Exponential	60	41	-1301.6	2980.9	126.2
Euclidean	Gaussian	5	7	-1558.7	3181.9	1.4
Euclidean	Gaussian	10	7	-1572.4	3209.2	1.4
Euclidean	Gaussian	15	8	-1572.4	3218.4	2.1
Euclidean	Gaussian	20	14	-1533.1	3195.1	2.3
Euclidean	Gaussian	25	15	-1522.3	3182.8	18.4
Euclidean	Gaussian	30	18	-1515.9	3197.5	16.4
Euclidean	Gaussian	35	28	-1490.3	3238.6	37.6
Euclidean	Gaussian	40	28	-1489.0	3235.9	38.3
Euclidean	Gaussian	45	34	-1479.6	3272.5	31.6
Euclidean	Gaussian	50	39	-1471.3	3301.9	23.8
Euclidean	Gaussian	55	47	-1452.3	3337.4	38.1
Euclidean	Gaussian	60	47	-1451.6	3336.2	103.9

Table S1. SALSA2D outputs for the Euclidean distance models

	Distance Type	Basis	Start Knots	End Knots	LogLik	BIC	Time (min)
25	Geodesic	Exponential	5	9	-1502.5	3087.8	2.2
26	Geodesic	Exponential	10	5	-1539.3	3124.7	2.7
27	Geodesic	Exponential	15	9	-1492.5	3067.9	4.1
28	Geodesic	Exponential	20	12	-1481.8	3074.1	5.1
29	Geodesic	Exponential	25	13	-1472.7	3065.2	21.5
30	Geodesic	Exponential	30	15	-1465.8	3069.8	18.6
31	Geodesic	Exponential	35	19	-1448.3	3071.6	40.7
32	Geodesic	Exponential	40	25	-1415.3	3060.9	13.5
33	Geodesic	Exponential	45	21	-1427.4	3048.2	45.8
34	Geodesic	Exponential	50	32	-1369.7	3034.2	62.7
35	Geodesic	Exponential	55	24	-1398.1	3017.3	90.7
36	Geodesic	Exponential	60	28	-1377.6	3013.1	103.7
37	Geodesic	Gaussian	5	8	-1551.0	3175.6	1.2
38	Geodesic	Gaussian	10	6	-1562.4	3180.1	0.9
39	Geodesic	Gaussian	15	7	-1553.0	3170.6	0.8
40	Geodesic	Gaussian	20	11	-1510.0	3121.4	18.5
41	Geodesic	Gaussian	25	14	-1500.4	3129.8	19.8
42	Geodesic	Gaussian	30	15	-1481.7	3101.6	13.9
43	Geodesic	Gaussian	35	21	-1472.5	3138.4	33.7
44	Geodesic	Gaussian	40	25	-1450.5	3131.3	8.7
45	Geodesic	Gaussian	45	26	-1433.8	3107.0	46.9
46	Geodesic	Gaussian	50	31	-1418.0	3121.7	40.9
47	Geodesic	Gaussian	55	32	-1408.3	3111.3	76.8
48	Geodesic	Gaussian	60	33	-1411.8	3127.5	26.6

Table S2. SALSA2D outputs for the Geodesic models

Distance Type	Basis	Start Knots	End Knots	LogLik	BIC	Time (min)
Euclidean	Exponential	60	41	-1301.6	2980.9	126.2
Euclidean	Exponential	55	40	-1318.0	3004.4	73.3
Geodesic	Exponential	50	32	-1369.7	3034.2	62.7
Euclidean	Exponential	50	32	-1374.5	3043.8	66.7
Euclidean	Exponential	45	28	-1376.8	3011.5	49.9
Geodesic	Exponential	60	28	-1377.6	3013.1	103.7
Geodesic	Exponential	55	24	-1398.1	3017.3	90.7
Euclidean	Exponential	40	27	-1407.7	3064.1	46.7
Geodesic	Gaussian	55	32	-1408.3	3111.3	76.8
Geodesic	Gaussian	60	33	-1411.8	3127.5	26.6

Table S3. Top 10 log-likelihood selected SALSA2D models

Distance Type	Basis	Start Knots	End Knots	LogLik	BIC	Time (min)
Euclidean	Exponential	60	41	-1301.6	2980.9	126.2
Euclidean	Exponential	55	40	-1318.0	3004.4	73.3
Euclidean	Exponential	45	28	-1376.8	3011.5	49.9
Geodesic	Exponential	60	28	-1377.6	3013.1	103.7
Geodesic	Exponential	55	24	-1398.1	3017.3	90.7
Geodesic	Exponential	50	32	-1369.7	3034.2	62.7
Euclidean	Exponential	50	32	-1374.5	3043.8	66.7
Geodesic	Exponential	45	21	-1427.4	3048.2	45.8
Geodesic	Exponential	40	25	-1415.3	3060.9	13.5
Euclidean	Exponential	40	27	-1407.7	3064.1	46.7

Table S4. Top 10 BIC selected SALSA2D models.

Model Averaged Selected Models

- k = number of knots
- r = effective radius sequence number
- w = AIC$_c$ model averaging weight

```
$expeuc
   k r w
1 50 1 1


$expgeo
    k  r           w
1  45  3 0.003343363
2  50  1 0.131329653
3  50  3 0.070500968
4  50  4 0.070848736
5  50  5 0.058365475
6  50  6 0.055646771
7  50  7 0.055112831
8  50  8 0.055009183
9  50  9 0.054987913
10 50 10 0.054925003
11 60  1 0.389930105


$gauseuc
   k r           w
1 55 2 0.006596805
2 55 3 0.052268277
3 55 4 0.062721223
4 55 5 0.025132387
5 60 2 0.107459248
6 60 3 0.566550328
```

```
7 60 4 0.159010615

8 60 5 0.020261116


$gausgeo

   k r          w

1 60 6 0.98712668

2 60 7 0.01287332
```

4. Full Analysis Covariate Information

Rainfall calculation

Fit a high dimensional smooth term to 156 locations of annual rainfall from 1999 to 2015 (2016/17 unavailable at the time of modelling) to interpolate values for the presence locations and pseudo-absence grid.

```
require(mgcv)
fit<-gam(meanrain ~ s(x.pos, y.pos,fx = TRUE, k=150), data=rainfall2)
analysisdat$meanrain<-predict(object = fit,
                    newdata = data.frame(x.pos = analysisdat$x.pos,
                                         y.pos = analysisdat$y.pos))
```

The results of the interpolation model are shown in Figure S5

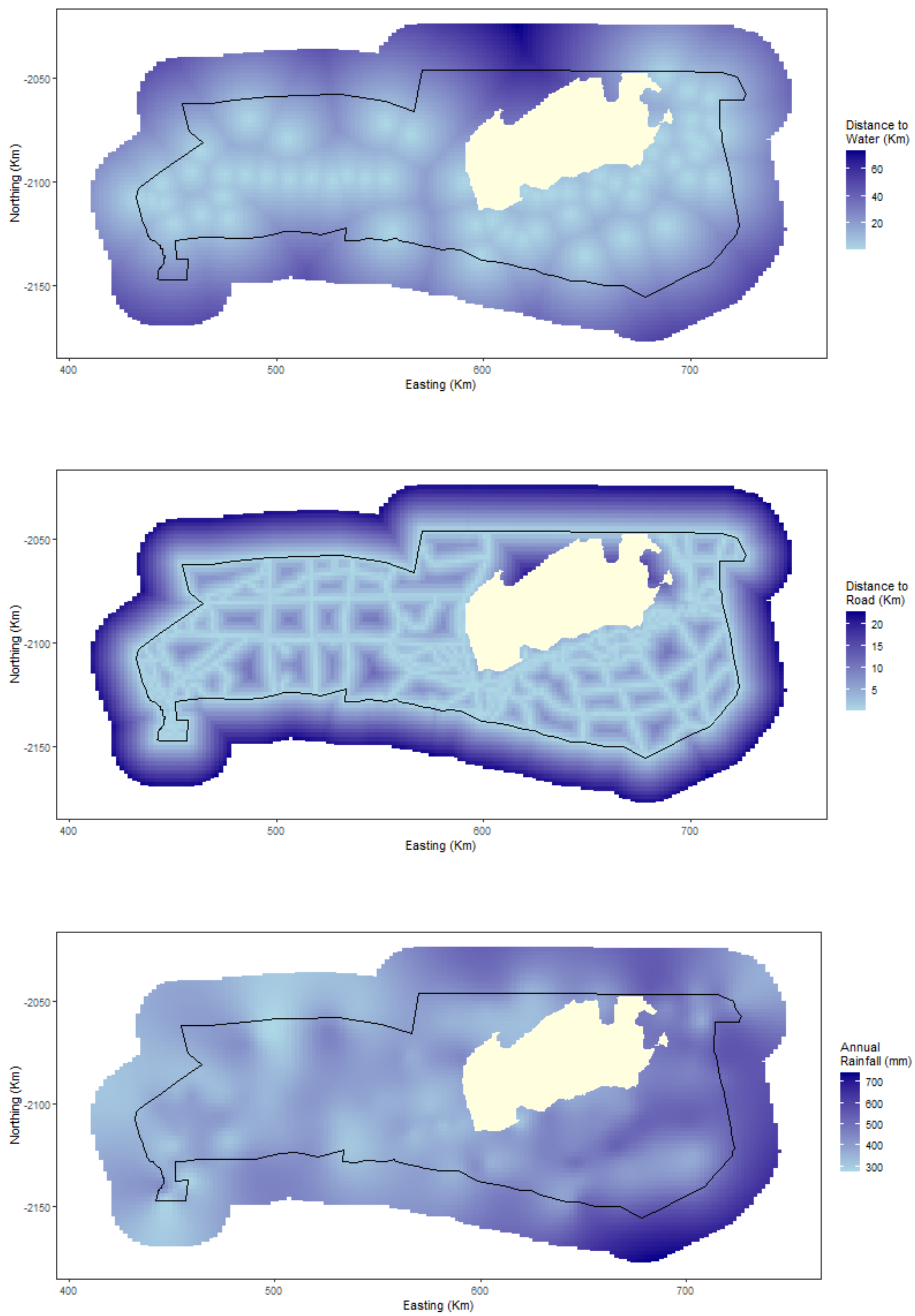

Figure S5. Figure showing the covariate data in the study region. Distance to water (top), distance to roads (centre) and interpolated annual rainfall (bottom).