Automated surface feature selection using SALSA2D: An illustration using Elephant Mortality data in Etosha National Park

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 June 24, 2022

Open Research Statement

The data and code are provided as private-for-peer review but can be made public if accepted for publication. The files can be found at the github site of the first/corresponding author: `https://github.com/lindesaysh/MIKE`. Additionally, the data file will be made public and permanently archived in the St Andrews PURE repository.

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

ABSTRACT

This analysis is motivated by the MIKE dataset in Etosha National Park (ENP). We use this dataset to show the development of an automated selection method for regression models to replace the model averaging used in the original CReSS paper. This method shows clear numerical and practical benefits over model averaging, and it's application to the elephant carcass data are of immediate and practical value to a range of stakeholders.

We have developed SALSA 2D in a GLM/GAM regression framework but this paper shows the flexibility of this approach by applying it to presence only data and use a downweighted Poisson regression. Using SALSA2D for model selection provided a more realistic local/clustered intensity surface compared with the model average approach.

The full analysis results showed high carcass intensity close to water holes and roads and in areas of the park with average rainfall. Some high risk areas were identified and these revelations are important for effective park management, particularly mitigation of poaching. It is impossible to patrol such a large area at random and these high intensity areas (particularly those accessed by a subset of roads and near some waterholes) can be targeted for more monitoring efforts than others.

KEYWORDS

CReSS (Complex Region Spatial Smoother); Spatially Adaptive; SALSA; spline; point process; GLM

1. Introduction

1.1. Motivating Example

The Monitoring the Illegal Killing of Elephants (MIKE; `https://cites.org/eng/prog/mike/index.php/portal`) programme is an international collaboration that collects and monitors trends related to the illegal killing of Elephants from across Africa and Asia [1]. The MIKE project also seeks to monitor the effectiveness of field conservation efforts and is part of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) initiative. MIKE operates in over 80 sites, across 43 elephant range states across Africa and Asia and rigorous protocols have been developed as part of this initiative to collect, analyse and build the capacity to better enforce the law and to reduce illegal elephant killings. The overall goal of MIKE is to provide information needed for elephant range states to make appropriate management and enforcement decisions, and to build institutional capacity within the range states for the long-term management of their elephant populations. In particular, they report PIKE (Proportion of Illegally Killed Elephants) for every range state.

The MIKE project has been active in Etosha National Park (ENP) in Namibia for over a decade, and substantial resources are used to collect relevant abundance and mortality data by dedicated aerial surveys under strict survey protocols. As part of routine park activities, opportunistic data also contributes to the African Elephant (*Loxodonta africana*) mortality database.

To meet the needs of the MIKE project and more generally, there is an urgent need to understand both the magnitude and spatial patterns of elephant deaths in ENP, regardless of cause. If the deaths are natural and, for instance, disease-related (e.g. anthrax) then this provides valuable information about the prevalence and locale of disease in the park. Endemic anthrax occurs in Etosha annually [2] and plays an important role in elephant population regulation/limitation. The monitoring of the prevalence of anthrax in elephant is important, because it advances our knowledge of a top down factor limiting a mega-herbivore. If instead, the deaths are a result of poaching then this provides necessary information about the prevalence, locale and patterns of these deaths.

Whilst in 2018 the poaching of elephants in ENP was low (20 deaths reported to MIKE and none poached), the general trend in more recent years is increasing (*pers. comm.* Etosha Ecological Institute). As the number of poached elephants increases it is very useful knowledge to have a baseline distribution of natural deaths. The development of statistical modelling methods aimed at predicting elephant mortality risk is crucial for early carcass detection. Should poaching occur in regions not common to find natural deaths, then increased/targeted mitigation measures can be efficiently actioned. Practically, understanding both the magnitude and spatial patterns of elephant deaths in ENP may assist in adapting patrol efforts in and around the park to track the anthrax disease and/or combat any poaching activities. It is also very important in light of the mass death events seen in Botswana in 2020 and 2021 [3]. Having a sense of “normal“ places of natural death may provide insights should such events ever occur in Etosha.

Statistical modelling of these data is necessary since the park is very large ($\sim$23,000 km^2) and regardless of the survey regime, the observed counts will undoubtedly comprise a subset of a larger number of deaths. Reliable modelling results which accurately estimate the magnitude and location of elephant mortality in ENP are also not guaranteed and require the careful consideration of at least the following two points, 1) most wildlife, including elephant, rarely traverse a large salt pan: there is little vegetation to be found in the pan and the sometimes boggy terrain prevents travel for large animals such as elephant; 2) the spatial patterns of mortality are likely to be localised and patchy: the abundance of elephant in the park is far from homogeneous and the reasons for death (natural or otherwise) are also likely to vary across the park. Failing to account for the possibly unusual spatial patterns in these data and/or assuming points across the pan are as closely linked as equidistant points without a physical barrier, can unwittingly lead to false conclusions about the magnitude and location of elephant deaths in the park.

The Complex Region Spatial Smoother (CReSS) is a regression spline based statistical modelling method equipped to address both aspects of these data [4]. Euclidean or geodesic (‘around the salt pan’) distances can be used to underpin the smoothed surface and the method is spatially adaptive enabling the targeting of surface flexibility to accommodate any particularly patchy trends and/or local surface features. While appropriate, the currently published CReSS method [4] undertakes the, crucially important, model selection process using a model-averaging of predictions approach which can be computationally intensive. We have also found after extensive use that this can mask unusually shaped spatial patterns when these are observed. In this paper, we propose using CReSS with an automated model selection approach, as an alternative to model-averaging, which enables atypical spatial patterns to be deduced from the data - patterns which have implications for park management in this case, and produces one model which is easier to handle.

1.2. Statistical development

The statistical development here involves the creation of a conceptually simple but effective heuristic algorithmic approach to carry out model selection for multidimensional basis functions to determine overall surface flexibility (via the number of 'knots') and the targeting of this flexibility (via 'knot' locations). The development of this algorithm was based on the Spatially Adaptive Local Smoothing Algorithm (SALSA) [5], which is for univariate smoothing. In order to distinguish the new algorithm, it will be referred to in this paper as SALSA2D.

We have also extended the suite of CReSS basis functions that can be used for the two-dimensional smoothing. This is useful since SALSA2D is agnostic about the basis function used but relies instead on an objective fit criteria for execution.

The new basis and SALSA2D algorithm are all implemented inside the `MRSea` `R` package [6, 7] for easy use by practitioners.

2. Methodology

2.1. The Complex Region Spatial Smoother(CReSS)

The published CReSS approach [4] achieves spatially adaptive surfaces via a judicious weighting of a variety of candidate surfaces with 'space-filled' knots [8] ranging from the very simplistic (via small numbers of knots with basis functions with a relatively global influence) to very complex (via large numbers of knots with basis functions with a relatively localised influence). This approach has also shown to perform well against other model-based alternatives developed for data sets with internal exclusion zones (such as coastlines and island systems) and is finding use in a range of ecological applications [9–11].

The CReSS approach fits pure spatial regression models to a set of coordinates $\mathbf{x}$ of the form:

$$g(\mathbf{y}) = \eta = \beta_0 + s(\mathbf{x}) \tag{1}$$

where g is the link function and η the linear predictor. $\mathbf{s}$ is a two dimensional surface approximated by a linear combination of exponential basis functions bE.

$$bE_{ki} = \exp^{(-h_{ki}/r_k^2)} \tag{2}$$

where r_k dictates the extent of the decay of this exponential function with distance between points, and thus the extent of its local nature. Notably h_{ki} indicates a geodesic or Euclidean distance (for some observation i and the k-th knot location). Parameter r_k takes values such that if r_k is small the model will have a set of relatively local basis functions and if r_k is large the model will have a set of relatively global basis functions. The exact values of r_k are dependent upon the range and units of the spatial covariates.

After the choice of distance metric, the CReSS with model averaging procedure fits multiple models with each model evaluated at one of a variety of parameter values for the number of knots (K) and the effective range parameter (r). According to Scott-Hayward et al. [4] model selection is achieved using AIC$_c$ [12] weights and averaging those models with ΔAIC$_c$ < 10 to produce weighted predictions.

While this approach (CReSS with model averaging) has been shown to produce reliable results in many cases [4], this procedure can be complicated, in terms of model handling and difficult to assess model fit and to provide confidence intervals. A paper by Dormann et al. [13] highlights some of the limitations of a model averaging approach. Namely, the authors show that estimating model weights introduces unknown and unaccounted for uncertainty and that confidence intervals for model-averaged predictions rarely achieve nominal coverage. They also state that model-averaging is most useful when the predictive error of contributing model predictions is dominated by variance (as opposed to bias), and if the covariance between models is low. We argue that ecological data, including the carcass data seen here, is often highly variable with limited covariates and thus could result in prediction errors dominated by variance. Additionally, given the CReSS with model averaging approach averages models with the same covariates but different parameterisations, there is also likely to be high covariance between competing models, rendering the model averaging approach less appropriate.

Further, when the spatial patterns are particularly unusual (e.g. stripe-like features or local hotspots are genuinely present) we have found that a model-averaging approach can result in overly smooth surfaces which mask these unusual, but important, patterns. This may result for a variety of reasons: under the original CReSS approach the space-filled knots are fixed in position for a given knot number, and the extent that each basis function is local (or global) is fixed (and the same) for all knots in that candidate surface.

As part of recent work, we have expanded the CReSS approach to include a Gaussian radial basis to the choice of basis functions available for selection. The two bases have different shapes, with the exponential being more peaked at the centre. These choices allow for more nuanced model fitting, akin to link function or distance metric choice. The Gaussian radial basis, bG, is specified as:

$$bG_{ki} = \exp^{(-(h_{ki}r_k)^2)} \tag{3}$$

where r_k and h_{ki} are as defined for the exponential (Equation 2) except that for the Gaussian basis, a small value for r_k returns a relatively global basis and a large r_k value returns a relatively local basis.

2.2. *Spatially Adaptive Local Smoothing Algorithm for at least two dimensions (SALSA2D)*

SALSA2D uses the same model framework as for model averaging (see Equation 1) but where k is chosen using an iterative three step procedure. The algorithm works in (at least) two dimensions and begins with space-filled knots to facilitate spatial coverage and then adaptively moves, adds and drops knots into, or from, locations in line with poor model fit (evidenced by large residuals) and an objective fit criteria. At each stage, the global/local extent of each basis function via the r_k value employed can also be revised as part of the search for a more appropriate surface. So, unlike the model averaging approach, SALSA2D returns one model with specifically selected k and r_k enabling standard methods for assessment of fit and uncertainty estimation.

The algorithm that drives SALSA2D has an iterative 3-step structure. After an initialisation step, there are three repeated steps: the first is a simplification step to reduce the number of estimated parameters which is achieved by allowing for the removal of columns from the design matrix (reduction in knot number). The second and third steps (exchange and improve) are designed to efficiently search the model space (all possible number and locations of knots). The exchange step allows for the possibility of moving away from a local optimum or addition of columns to the design matrix (a new knot) and the improvement step attempts to make local improvements in knot location. The outcome of each of these steps is determined by an objective fit criterion and repeated until no improvements are made (or an iteration limit is reached). The structure of the algorithm is given in the pseudo code in Figure 2.2 and the next sections describe the steps in detail.

2.2.1. Initialisation

Each observed location, i, is considered a possible location for a knot position. To avoid estimation issues, only unique knot locations are considered giving K_l legal knot locations. The user specifies a starting number of knots, K_s, where $K_s < K_l$, and these are selected from K_l using a space-filling algorithm [8]. This method pro-

SALSA2D:
Given an n-dimensional set K_l of possible knot locations over the region of interest.
    Initialise
        Initialise knots, K_s within the points of K_l
        Check for convergence
    Repeat
        Repeat Simplification step *while* ($K > K_{\text{min}}$ and fit measure improves)
        Repeat Exchange step *while* ($K < K_{\text{max}}$ and fit measure improves)
        Repeat Improvement step *while* (fit measure improves)
    While (an improvement in fit measure is made by one of the above steps)

Figure 1. Pseudo-code outlining the structure of SALSA2D [adapted from Figure 1, 5], where K is the number of knots used for fitting.

vides good coverage across the spatial region as a starting position for SALSA2D. Additionally, the minimum number of knots, K_{min} ($2 \leq K_{\text{min}} < K_s$) and maximum number K_{max} ($K_s < K_{\text{max}} \leq K_l$) are specified.

To evaluate the basis function, the r_k-value for each basis must also be chosen. The SALSA2D algorithm selects from R possible options for r_k which range from a very local basis to a globally acting basis. The middle option which is neither very local or very global, is chosen to initialise the first model.

To ensure that the initial model fit has converged, there is a drop step component that is activated if the variance of the initialised first model exceeds that of the simpler input model (the variance should not increase with additional parameters/flexibility in the model). If this occurs, knot locations with the largest contributions to the variance are removed one by one until the overall variance of the more complex model is lower than the input model.

2.2.2. The simplify step

Using the fit criteria specified, the simplify step compares the current model with all models obtained by removing an existing knot (as long as this is at least K_{min}). At each iteration, the model with the best fitness measure is retained and the process repeated until there is no further improvement in the fitness measure. This step can be carried out by fixing r_k or by choosing r_k for each basis as each knot is dropped for comparison.

2.2.3. The exchange step

The exchange step increases the extent of the search of model space by enabling a move away from a local minima (of the fit criterion). It uses the maximum Pearson residual from the current fitted model to identify a possible candidate location for a new knot (although in theory other types of residuals could be chosen and we use an alternative metric for the point process models in the next sections). The algorithm then compares the objective fit criteria for these models that result when each of the existing knots in the current model is moved to this new location, and also the fit criteria from the model that results when an additional knot at this location is added to the current model (if this does not exceed $K_{\max}$). The model with the best fitness measure is retained in this step if it has a better fitness measure than the current model. Evaluation of each of these models can be very quick to return but this process is naturally more computationally expensive, if r_k is also chosen for each basis function for each candidate model. In practice, the algorithm uses the knot locations of the five largest residuals as candidates for an exchange or move.

2.2.4. The improve step

The improve steps allows a more nuanced search of the local minima by allowing small adjustments to the location of each knot. Using the fit criteria specified, the improve step compares the current model with all models obtained by moving an existing knot to one of its five nearest neighbours (determined by the distance metric employed: geodesic or Euclidean). At each iteration, the model with the best fitness measure is retained. As with the exchange step, alternative choices for the r_k parameter may be considered when fitting each new model and this process is likely to be swift at this stage.

2.2.5. Determining r_k

This routine considers incrementing or decrementing r_k values in the sequence of R possible values, where the sequence is selected using the method from 4. It can be evaluated either once at the end of the exchange, improve and simplify steps or as part of every decision taken during these steps. The process is done by considering each of the radial basis columns in turn, and incrementing or decrementing the r_k values in the index until there is no improvement in the fitness measure. At each step the r_k-values for the other basis columns are maintained at the current solution. The best of these models is selected as the new current model, and the process iterates until no improvement is made. This process can have a large computational overhead and may significantly prolong the procedure but constitutes a broader search of the model space.

This algorithm is implemented in the `MRSea` package which can be found at `www.github.com/lindesaysh/MRSea` [6].

3. Methods comparison

This section compares the performance of the CReSS with model-averaging approach to CReSS with SALSA2D for model selection. The methods are compared numerically using log-likelihood, while the practical consequences of using each are assessed visually and contextualised using surface features in Etosha National Park, Namibia.

3.1. Data specification

To appreciate the numerical and practical benefits of this methodological development, the MIKE data was used for the method comparison and for the subsequent analysis in full [1]. These data consists of 320 carcass locations observed between February 2000 and March 2017 in Etosha National Park (ENP). The observed fatalities are recorded as being due to: anthrax, natural (age-related) causes, poaching and unknown. While a substantial proportion of the carcasses are recorded as being for ‘unknown’ reasons (54%) the largest known cause of death is from Anthrax (27.8%). Less than 1% of the carcasses were confirmed as poached. Disregarding 2017 as it was only a partial year, 2006, 2013 and 2014 had the fewest recorded carcasses (7-8), whilst 2002, 2003, 2005 and 2011 had the highest recorded (27-28). As there were a relatively small number of observations per year, no guarantee the deaths occurred in the year of detection and no obvious changes in the spatial pattern of observations, the data were pooled across years.

The longitude and latitude coordinates were converted to Universal Transverse Mercator (UTM) zone 33S and the study region was extended beyond the ENP boundary by 20 km to allow for the inclusion of carcasses just outside the park. Additionally, the large salt pan was reduced in size by 2 km to allow inclusion of carcasses found near the edge of the pan. The data show that carcasses generally seem to occur near roads (or, at least, are more commonly observed near roads) and waterholes (Figure 2). It is possible that these patterns are due to opportunistic reporting of carcasses as a result of park vehicles moving along the roads, however the data were from both opportunistic and dedicated surveys, which are carried out without reference to roads. Furthermore, collared elephant in ENP have been shown to utilise roads/tracks and fire breaks extensively and are known to frequent waterholes [14].

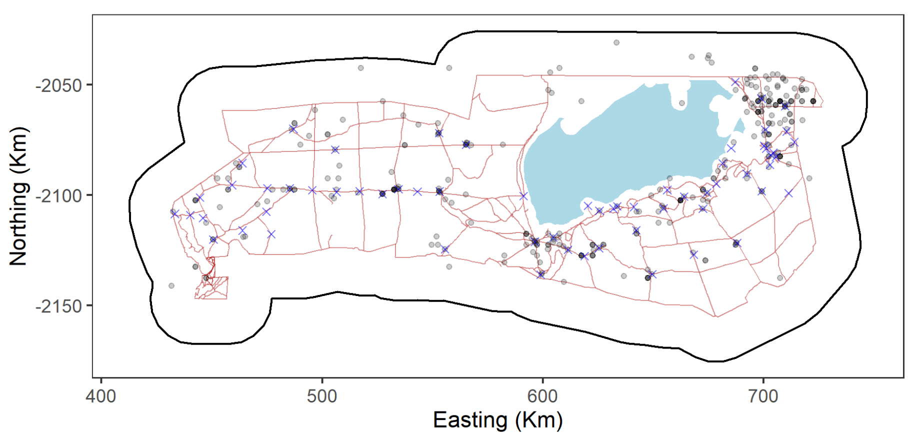


Figure 2. Figure showing the study area (ENP) with the carcass locations shown as dots. As there are duplicate locations, the darker the dot, the more presence locations. The study area (park boundary plus 20km buffer) is outlined in black. The blue polygon is the Etosha salt pan, the red lines are park roads and the blue crosses are waterholes. The outermost red line is also the park fence.

In much of the grey literature the methods described here have been applied in a Poisson or Binomial generalised additive model framework (GAM). Here we have chosen to showcase the versatility of the SALSA algorithms and apply them to a presence only data set, where the primary interest are the spatial locations of presence points (carcass locations). In this data set, the link back to the original survey effort (where surveys were undertaken) is not available so we are left with only the carcass locations and no absence locations. Warton and Shepherd [15] showed the link between logistic regression and an inhomogeneous Poisson point process model (PPM) and here we use both this link and the downweighted Poisson regression method [16] to fit a Poisson PPM using a pure regression GAM framework. In this case, the intensity is the number of presence records (carcass sightings) per unit area and is modelled as a function of covariates measured throughout the study region. It is a relative measure and gives the expected abundance of carcass sightings for a given area.

Pseudo-absences in the regression setting play the same role as quadrature points in point process modelling and we used the point process framework to choose the number and location of these points. The pseudo-absence points were selected as a regular grid and the number based on convergence of the likelihood [16].

Lastly, to determine areas of poor fit, the exchange step requires the calculation of residuals. This was achieved by creating a neighbourhood around each knot location (k) and comparing the observed number of points with the sum of the estimated intensities in the same area. For more details, see Section 1 of the Supplementary Material.

3.2. Model specification

To compare the performance of SALSA2D with model averaging as a model selection approach, models with a two dimensional smoother-based term for geographic locations were fitted to the MIKE data. The comparison involved either the published CReSS method which employs model averaging [4] or model selection using SALSA2D to determine knot number and location. Here we model the the locations of the carcasses jointly with the pseudo-absences by maximising the following weighted Poisson log-pseudolikelihood [17]:

$$l(\beta; \mathbf{X}) = \sum_{i=1}^{N} w_i (y_i \log(\lambda(\mathbf{X}_i)) - \lambda(\mathbf{X}_i)) \tag{4}$$

where $\lambda(\mathbf{X}_i))$ is the intensity at location i, $\boldsymbol{X}_i$ represents the design matrix at location i, N is the total number of points (presence and pseudo-absence), $\mathbf{w} = \{w_1, \cdots, w_N\}$ are quadrature weights.

$$y_i = \begin{cases} \frac{1}{w_i} & \text{if } i \text{ is a presence location} \\ 0 & \text{if } i \text{ is a pseudo-absence location} \end{cases}$$

The log-pseudolikelihood in Equation 4 [17] is a re-expression of the Poisson PPM log-likelihood [18], which means that models can be fitted using standard GLM software. Here we model the expected number of carcasses per square kilometre and so the weights for the pseudo-absence points are specified as the area of the study region, 37,872 km^2 (ENP plus the 20 km buffer) divided by the number of pseudo-absences. The weights for presence points are set to some small value (10^{-6}). Likelihood convergence was used to determine the the number of pseudo-absences which was estimated to be 9644 (a grid spacing of 2 km). For more details see Section 2 of the Supplementary Material.

For this method comparison section, we model the intensity as a function of coordinates, $\mathbf{x}$, only.

$$\log(\lambda(\mathbf{X}_i)) = \eta_i = \beta_0 + s(\mathbf{x}) = \boldsymbol{X}_i^T \boldsymbol{\beta} \tag{5}$$

where η_i is the linear predictor, consisting of the intercept, β_0, and a smooth function of coordinates, s($\mathbf{x}$). The smooth function is either the exponential or Gaussian basis function.

For both the model averaging and SALSA2D methods, the following specifications were used to return the columns of the design matrix $\boldsymbol{X}$ in Equation 5:

- Two basis options: Exponential (bE_{ki}; Equation 2) or Gaussian (bG_{ki}; Equation 3)
- Two distance measures (Euclidean or geodesic) to calculate h in the basis equations; the geodesic distances are calculated using Floyds algorithm [19] and for more details see [4].
- 12 choices of fixed knot number (for the model-averaging approach) and 12 choices of starting knot numbers, K_s for the SALSA2D approach. In each case, the fixed/starting knot set was: [5, 10, 15, . . . , 55, 60]. A total of 285 legal knot positions (K_l) were considered. These consisted of all non-duplicated carcass locations (n=245) and 50 space-filled pseudo-absence locations ($\sim$ 20% of all K_l).
- 10 choices of r_k (also specified as part of Equations 2 and 3)

Additionally, for SALSA2D, $K_{\min}$ and $K_{\max}$ were set to 2 and 100 respectively, for all model specifications.

3.3. Model comparison

In keeping with Scott-Hayward et al. [6], the model-averaging CReSS method was governed by AIC$_c$ weights which were used to choose which models to average ($\Delta AIC_c \leq 10$) and their relative contribution to the overall averaged model. In keeping with Walker et al. [5], the BIC was used to govern SALSA2D model selection regarding the choice of knot number and their locations across the range of combinations of basis type, distance metric, starting knot number and r_k choices [20]. In all cases, the log-likelihood score (Equation 4) was calculated for each model to enable comparison between model selection strategies.

4. Results

4.1. Numerical comparison

The log-likelihood scores returned for the model averaging method were fairly close (maximum difference 14 points) regardless of the basis function and distance metric used in each model (Table 1, Method: 'Model averaging'). The geodesic-exponential combination scored the best (largest log-likelihood) of the 4 combinations trialled. Interestingly, this combination chose 11 models with which to average over to obtain this solution, compared with some options that chose far fewer models to use as part of the average calculation. In general, geodesic distances were preferred to Euclidean regardless of basis.

Table 1. Table showing the results of the model averaging and SALSA2D methods of model selection for a given basis type and distance metric used. The 'No. Models' indicates the number of models chosen to carry out the model averaging in each case, and the 'No. Knots' indicates the number of knots chosen for each model using the SALSA2D selection method. The star indicates the model with the largest log-likelihood (LL) score, and thus the chosen model based on the LL in each case.

Method	Basis	Distance Measure	No. Models	No. Knots	Log-Likelihood
MA	Exponential*	Geodesic	11	-	-1432.0
	Gaussian	Geodesic	2	-	-1441.5
	Exponential	Euclidean	1	-	-1443.4
	Gaussian	Euclidean	8	-	-1446.3
SALSA2D	Exponential	Geodesic	-	32	-1369.7
	Gaussian	Geodesic	-	32	-1408.3
	Exponential*	Euclidean	-	41	-1301.6
	Gaussian	Euclidean		47	-1541.6

The log-likelihood scores for the SALSA2D based selection are shown for the model with the highest log-likelihood for each of the basis/distance metric combinations

(Table 1, Method: SALSA2D). Across the four combinations, the scores were less homogeneous than for the model averaging results and the exponential-Euclidean SALSA2D model (using 41 knots) was the best of all trialled here. In contrast to the averaging approach, there was a preference for the exponential basis with the distance metric secondary. In reality, the user may prefer to select the best model using BIC (as was used for k/r selection). In this case, the order of the four parameterisations was the same (exponential-Euclidean the best and Gaussian-Euclidean the worst) and the best model using BIC was the same as in Table 1 when log-likelihood was used (see Section 3 of the supplementary material for an expanded version of Table 1).

Using the "best" SALSA2D models only, for all but one combination of basis type and distance metric used, all SALSA2D models produced better scores than the model averaging method – sometimes reducing the log-likelihood score by as much as 10%. However, if SALSA2D initialises with too few knots, the algorithm may get stuck in local minima. So long as a large enough number of starting knot locations was selected ($\sim\geq$ 40), SALSA2D-based selection resulted in superior scores over the model-averaging alternative (Figure 3). This demonstrates that the SALSA2D model selection method can return improved results and at worst, SALSA2D results were almost indistinguishable from the best model averaging-based result.

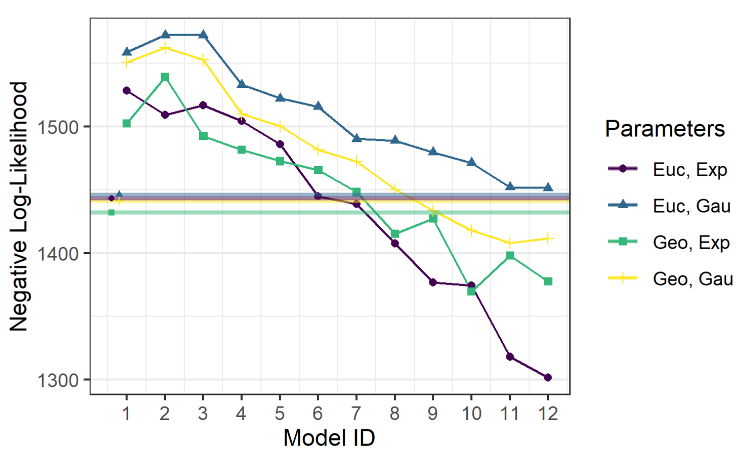


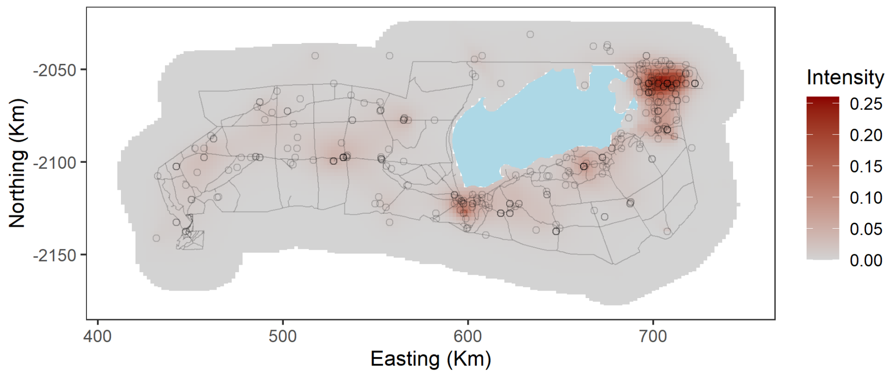
Figure 3. Figure showing the model identification number (increasing start knots) and the negative log-likelihood score for each of the SALSA2D models resulting from a different start knot number, K_s. The horizontal lines are the scores for the equivalent model averaging result.

4.2. Visual comparison

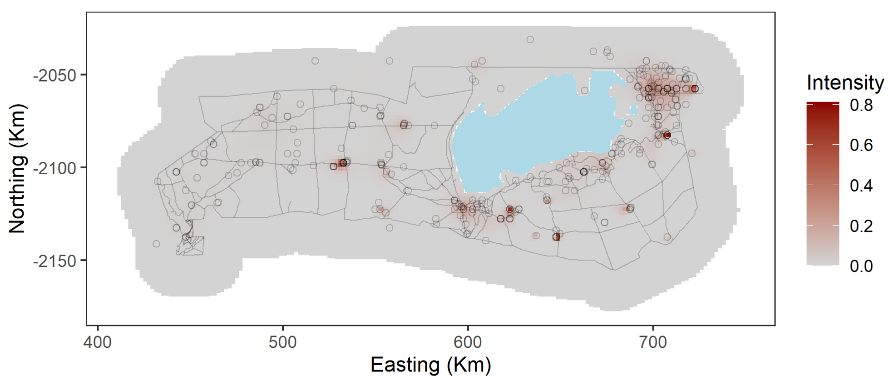
Results for the model-averaging based model (Figure 4a) signalled that the intensity of carcasses was highest in the north-east of the park (where most the observed deaths occurred) and along the southern edge of the large salt pan, which is consistent with the observed data. The carcass intensity is very low near the south-west and South-Eastern borders of ENP.

Whilst the model averaging results show a smooth intensity surface, the SALSA2D method produces a more clustered intensity surface (Figure 4b). The surface shows more local effects, particularly the centre west and below the salt pan and the highest intensity at these spots was nearly three times that of the model-averaging result. These effects, match well with the carcass location data and frequently occur at the confluence of several roads and some waterholes.

Figure 5 shows the selected knot locations and equivalent r parameter from the 11 averaged models (Figure 5a) and the one best SALSA2D model (Figure 5b). The averaged knot locations are more difficult to represent but it can be seen that there are multiple r values (ranging from global to very local) across the same locations and occasionally a location where the sign of the coefficient changes between models. The SALSA2D result is more nuanced with very few knot locations selected to the west of the park. For the 41 selected locations, a variety of r's were chosen. It is interesting that the SALSA2D approach found the Euclidean distance metric to be best and it is possible that the more local knots chosen under this method negate the need for the geodesic distances by limiting the possible leakage across the pan.

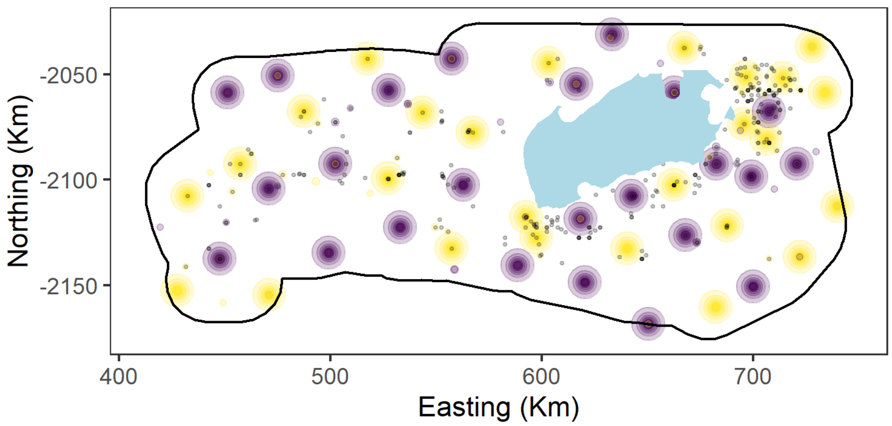


(a) Model Averaging

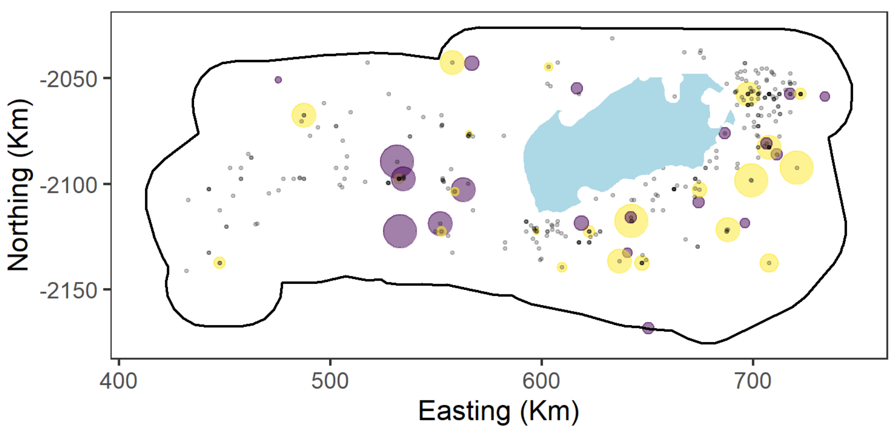


(b) SALSA 2D

Figure 4. Figure showing the intensity of carcass locations throughout the study area (ENP) using the model averaging (top) and SALSA2D (bottom). Note: to ensure detail can be seen, the two images have differing intensity scales. The carcass locations are shown as black circles. The blue polygon is the Etosha salt pan and the lines are the roads within the park.



(a) Model Averaging (11 models)



(b) SALSA 2D

Figure 5. Figure showing the knot locations and r from the best model averaging (top) and SALSA2D (bottom) models. Yellow is for a positive model coefficient and purple a negative one. The size of the coloured circles is a visual representation of the size of the r parameter. Note that in (a) the concentric rings are from models had the same knot locations with different r. In (b) the colours overlap but each k is in a different location. The carcass locations are shown as grey/black circles. The blue polygon is the Etosha salt pan.

5. Application: Analysis of Elephant mortality in Etosha National Park

5.1. Data available

The intensity of elephant carcasses, based on the observed carcass locations and pseudo-absences, was modelled using four candidate covariate terms: distance from the nearest road, distance from the nearest water point, mean annual rainfall and a spatial term based on spatial coordinates (in km, UTM zone 33S).

The distance from nearest road and nearest waterhole metrics were calculated using shape files supplied by the Ministry of Environment and Tourism (Namibia). These metrics were considered as candidates in the model to reflect possibly differential mortality rates near roads and waterholes, regardless of their spatial location in the park.

The mean annual rainfall was based on rainfall data collected from 168 rain gauges distributed across Etosha National Park which are visited annually, when possible. Annual rainfall was not available for every gauge for every year, due to logistical difficulties reaching remote areas in some years, and so this metric was averaged across years for each gauge before interpolation to indicate areas in ENP with persistently high or low rainfall. The interpolation was achieved using a high dimensional penalised spline ($df = 150$) to allow for interpolation to the carcass data locations and to the pseudo-absence grid. Details on the rainfall interpolation can be found in Section 4 of the supplementary material.

The proximity to waterholes was included as a candidate since elephant frequent water holes throughout the year, particularly in the dry season; roughly May to October [14] and have been shown to have increased habitat use with proximity to water. [21].

While natural deaths might occur in line with their distributional patterns it is thought Anthrax-related deaths may be related to the use of water holes [22]. The relationship with waterholes was found to be very stepped and so this variable was converted to a 2 level factor; < 3km and ≥ 3km (cutoffs of 1-5km were trialled and assessed using BIC).

The reasons for including proximity to roads as a candidate might seem less obvious, but the attraction or repulsion to roads by elephants might also be evident in their mortality patterns, and the model comparison work demonstrated that some roads are important (Figure 4b). This could be due to elephant preference to be found near roads, which is possible owing to their extensive use of roads/tracks for travel [14], but can only be confirmed by a dedicated analysis of survey data or that the detection of carcasses is higher near roads (e.g. easier to observe).

The spatial term was considered for inclusion in this model to represent the spatial patterns in mortality that are not adequately explained by proximity to roads, water holes or annual mean rainfall. The role of this term in this model is crucial in this case - correctly identifying systematic spatial patterns in mortality might provide insights about other park features not currently considered to be related to mortality and overlooking these features prevents the mitigation of future elephant mortalities, particularly those related to poaching.

5.2. Model specification

We are interested in modelling the intensity of elephant carcass locations as a function of distance to water, roads, mean annual rainfall and as a spatially adaptive smooth function of spatial coordinates. The model specification was:

$$\begin{aligned}\log(\lambda(\mathbf{X}_i)) &= \eta_i \\ &= \beta_0 + \text{distWater}_i + s_1(\text{rainfall}_i) \\ &\quad + s_2(\text{distRoads}_i) + s_3(\mathbf{x}) \\ &= \boldsymbol{X}_i^T \boldsymbol{\beta}\end{aligned}$$

In this case, $\lambda(\mathbf{X}_i)$ is the intensity at location i and $\mathbf{X}_i$ represents the coordinates and environmental covariates. s_1 and s_2 represent one-dimensional basis functions, while $s_3(\mathbf{x})$ represents a two-dimensional exponential basis function for the spatial coordinates. $\boldsymbol{\beta}$ is a vector of model parameters associated with all columns of the design matrix, $\mathbf{X}$. The columns of $\mathbf{X}$ comprise the intercept (1), water $\geq$ 3km (0,1), B-spline bases for rainfall and roads and the exponential radial basis for the spatial term.

Specifically, quadratic B-splines with SALSA based knot selection [5] were used to implement the one dimensional smooth terms for rainfall and roads. The two-dimensional spline basis function was determined using Equation 2 (exponential basis) and based on Euclidean distances. Knot number, their locations and r_k values were chosen using the SALSA2D algorithm. The starting parameters were based on the best result from the simulation study; $k_s = 41$, $k_{\min} = 2$ and $k_{\max} = 100$. The BIC was used to govern model selection in all cases.

5.3. Results

The results show that carcass intensity is highest near to water holes and roads (Figures 6 & 7) and locations where the annual rainfall is approximately 450mm (Figure 8). Specifically, intensity decreases steeply with the distance from road until approximately 1km when the relationship subsides.

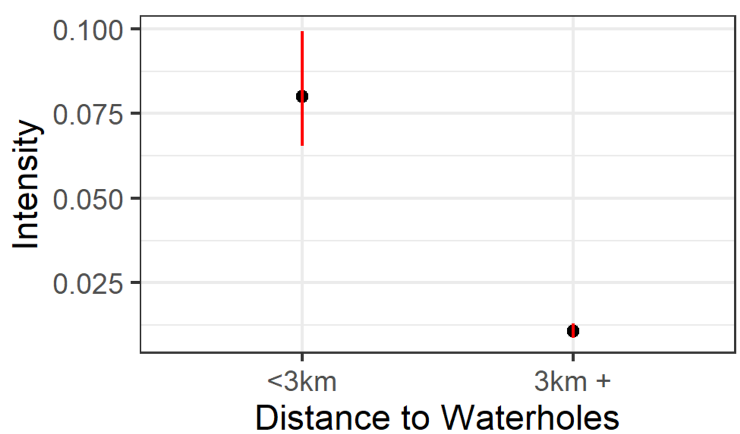


Figure 6. Figure showing the estimated relationship of distance to the nearest waterhole to carcass intensity (when distance to roads = 1.5km and mean annual rainfall = 420mm). The red line area is a 95% confidence interval about the estimated relationship.

The addition of distance from roads and mean annual rainfall to the spatial term, improved model results when compared with model results based on a SALSA2D-based spatial term alone (Models 2 and 3 Table 2); the BIC scores substantially im-

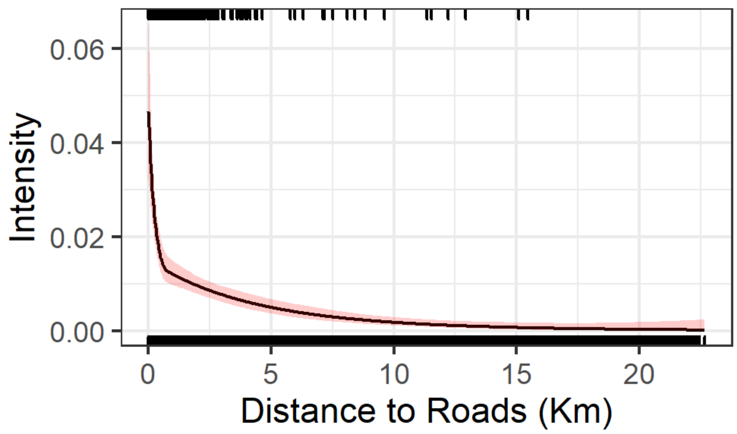


Figure 7. Figure showing the estimated relationship of distance to roads to carcass intensity (when mean annual rainfall = 420mm and distance to waterhole is $\geq$ 3km). The red shaded area is a 95% confidence interval about the estimated relationship. The tick marks top and bottom show the values of the covariate in the original data which were presence locations (1's) and absences (0's).

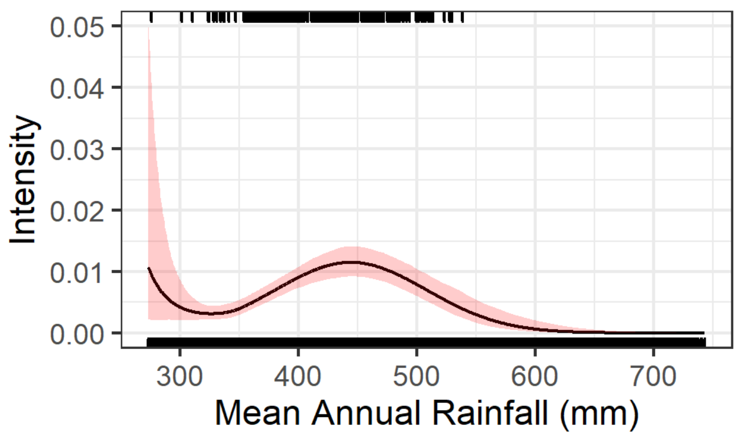


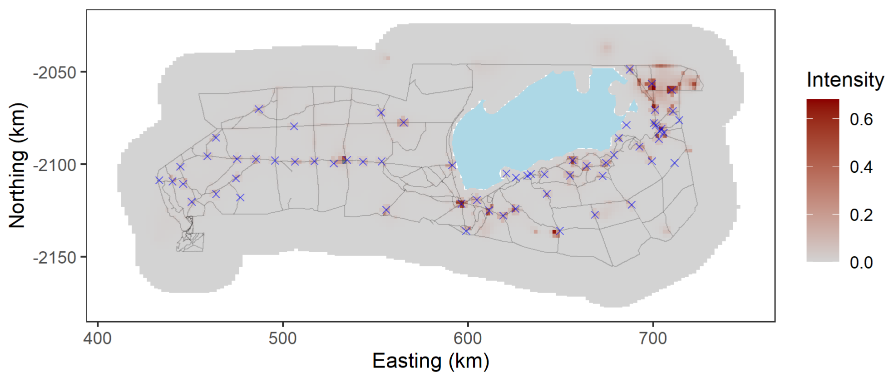
Figure 8. Figure showing the estimated relationship of mean annual rainfall to carcass intensity (when distance to road = 1.5km and distance to waterhole is $\geq$ 3km). The red shaded area is a 95% confidence interval about the estimated relationship. The tick marks top and bottom show the values of the covariate in the original data which were presences (1's) and absences (0's).

proved from 2980 to 2848.

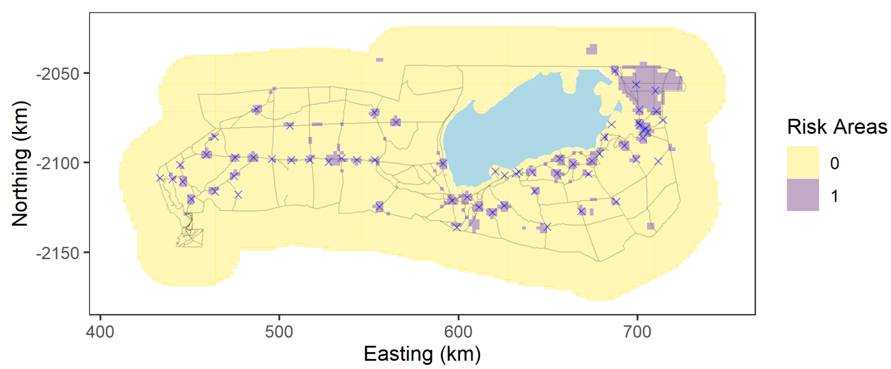
The spatial term also contributed positively to the model, despite the extra parameters incurred (Table 2); the BIC score decreased from 3084 for the univariate model (Model 1) to 2848 when the spatial term was included (Model 2). The practical consequences of its inclusion was clearly evidenced by tempering the 'global' effect of roads and water which was implicit in the model that included the additional variables (Figure 9a). In some cases the road and water effects diminished altogether where carcasses were not seen in the data. Crucially, this spatial term also better accommodates carcass locations which are not explained by only their proximity to water, distance to roads or average annual rainfall. Figure 10 shows that in Model 1, the water hole relationship dominates with a peak of intensity at each one. When the spatial term is added, the waterhole peak is suppressed at a number of waterholes and even increased at others. The peak in intensity is shifted to the north which is in keeping with the high number of carcasses observed there. The knot locations are similar to Model 3 but with fewer in the west and a higher proportion of smaller r (Figure 10b). Overall, the modelling shows that most, but not all, waterholes and some roads have high carcass intensity. Figure 9b shows the top 5% highest carcass intensity areas which form the highest risk areas in the park.

Table 2. Table showing the results for the model based on one dimensional smoother-based relationships only (model 1) and the model with both one and two dimensional smoothers (model 2). For reference, model 3 is the model with only a two dimensional smooth (see Table 1).

Model	Term	*df*	χ^2 p-value	Log-Likelihood	BIC
1	s(rainfall)	3	$p < 0.0001$	-1505.4	3084.4
	s(distRoads)	3	$p < 0.0001$		
	Near water	1	$p < 0.0001$		
2	s(rainfall)	3	$p < 0.0001$	-1221.3	2848.0
	s(distRoads)	3	$p < 0.0001$		
	Near water	1	$p < 0.0001$		
	s(xcoord, ycoord)	36	$p < 0.0001$		
3	s(xcoord, ycoord)	41	$p < 0.0001$	-1301.6	2980.1

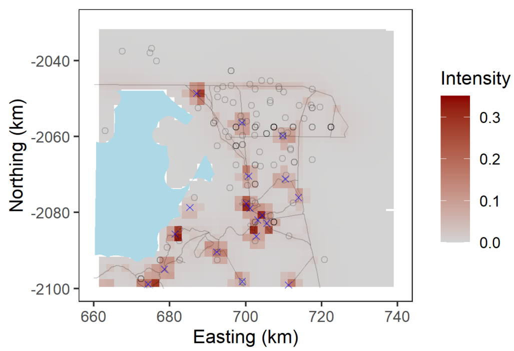


(a)

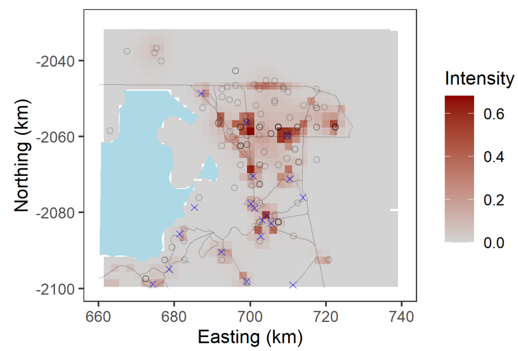


(b)

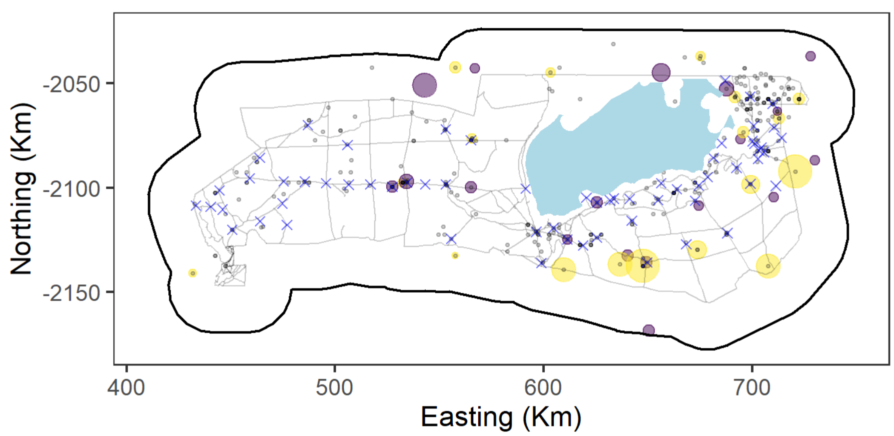
Figure 9. Figure showing the estimated carcass intensity throughout the study area using SALSA and SALSA2D-based model selection and both one and two dimensional spline based terms (a). Figure showing the top 5% intensity areas. The carcass locations are shown as black circles, the blue polygon is the Etosha salt pan, the blue crosses are waterholes and the black lines are roads.



(a) Model 1



(b) Model 2



(c) Yellow dots = positive coefficient, purple dots = negative coefficient. The size of the coloured dots is proportional to the size of r

Figure 10. Figure showing the estimated carcass intensity for the north east part of the study area for (a) Model 1, (b) Model 2 and (c) the location of k and associated r for model 2. The carcass locations are shown as black circles, the blue polygon is the Etosha salt pan, the blue crosses are waterholes and the black lines are roads

6. Discussion

Using SALSA2D for model selection provided better results and the ability to have a more realistic local/clustered intensity surface compared with the model averaging approach. There are clear numerical and practical benefits to SALSA2D-based model selection compared with a model averaging approach in this case and while the benefits of doing so might be less stark in cases where spatial patterns are more smooth, it needs to be possible to identify clusters and irregular patterns, such as those observed here, when they exist.

Simply including proximity to water and roads in the model as part of this analysis did not reveal genuine patterns in all areas of the park, since not all roads/waterholes have been associated with carcasses. The addition of the spatial term with spatially adaptive knot selection was able to suppress/enhance the global relationships with the environmental covariates in particular areas. This resulted in the identification of some critical areas of the park which is important for effective park management, both in terms of disease outbreak, which after 'unknown' was the largest category, and poaching, which although small in this data set (<1% of carcasses), occurs in the park. It is impossible to patrol such a large area at random and the areas of the park identified here (particularly those accessed by a subset of roads/waterholes) appear to require more monitoring efforts than others. Elephants are highly mobile and so early detection of carcasses, in particular anthrax related deaths, are important to identify spread of disease across the park [23].

This extended CReSS approach using SALSA2D model selection is of immediate and practical value to a wide range of users of statistical modelling methods. SALSA2D is implemented inside the `MRSea` package and it can automatically select knots based on two user-defined types of two-dimensional spline bases (Gaussian and exponential) and distance calculation (Euclidean or geodesic) based on a range of objective fitness criteria, chosen by the user. Notably, exclusion zones and non-Euclidean distances can be included to model more complex spatial regions [as seen in 4, 6] and adaptations have been made to allow for the fitting of Poisson PPMs using the downweighted Poisson regression method. By using presence only data in this paper, as opposed to the more traditional Binomial or Poisson GAMs, we have demonstrated the flexibility of this approach for a wide variety of settings.

7. Supplementary Material

See additional document for information on residual calculation, pseudo-absence selection, expanded results of the methods comparison and details of rainfall interpolation.

Acknowledgements

The authors would like to thank Dr. Richard Glennie for his assistance and patience with PPM queries and reading a draft of the paper.

Notes on contributors

- LSH, MLM and CGW contributed to method development, analysis and paper writing
- GS, WK and PdP contributed to data collection and local information

Disclosure Statement

There are no competing interests.

Funding

The authors declared that no grants were involved in supporting this work.

References

[1] MIKE. https://cites.org/eng/prog/mike, 2018. URL `https://cites.org/eng/prog/mike`.

[2] W. C. Turner, P. Imologhome, Z. Havarua, G. P. Kaaya, J. K. E. Mfune, I. D. T.

Mpofu, and W. M. Getz. Soil ingestion, nutrition and the seasonality of anthrax in herbivores of Etosha National Park. *Ecosphere*, 4(1):13, 2013. .

[3] T. Karombo. Elephants are dying in droves in botswana. scientists don't know why, 2021. URL `https://www.sciencenews.org/article/african-elephant-mass-death-botswana`.

[4] L. A. S. Scott-Hayward, M. L. Mackenzie, C. R. Donovan, C. G. Walker, and E. Ashe. Complex Region Spatial Smoother (CReSS). *Journal of Computational and Graphical Statistics*, 23(2):340–360, 2014.

[5] C.G. Walker, M.L. Mackenzie, C.R. Donovan, and M.J. O'Sullivan. SALSA - a Spatially Adaptive Local Smoothing Algorithm. *Journal of Statistical Computation and Simulation*, 81(2):179–191, 2010.

[6] L.A.S. Scott-Hayward, Mackenzie M.L., and C.G. Walker. MRSea package: Statistical modelling of bird and cetacean distributions in offshore renewables development areas, 2021.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

[8] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

[9] D. J. F. Russell, G. D. Hastie, D. Thompson, V. M. Janik, P. S. Hammond, L. A. S. Scott-Hayward, J. Matthiopoulos, E. L. Jones, and B. J. McConnell. Avoidance of wind farms by harbour seals is limited to pile driving activities. *Journal of Applied Ecology*, 53(6):1642–1652, 2016.

[10] R. A. Dunlop, M. J. Noad, R. D. McCauley, L. Scott-Hayward, E. Kniest, R. Slade, D. Paton, and D. H. Cato. Determining the behavioural dose–response relationship of marine mammals to air gun noise and source proximity. *Journal of Experimental Biology*, 220(16):2878–2886, 2017.

[11] D. V. Harris, J. L. Miksis-Olds, J. A. Vernon, and L. Thomas. Fin whale density and distribution estimation using acoustic bearings derived from sparse arrays. *The Journal of the Acoustical Society of America*, 143(5):2980–2993, 2018.

[12] Nariaki S. Further analysts of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.

[13] C. F. Dormann, J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn,

K. Bartoń, C. M. Beale, S. Ciuti, J. Elith, K. Gerstner, et al. Model averaging in ecology: a review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 2018.

[14] M. Tsalyuk, W. Kilian, B. Reineking, and W. M. Getz. Temporal variation in resource selection of African elephants follows long-term variability in resource availability. *Ecological Monographs*, 89(2):e01348, 2019. .

[15] D. I. Warton and L. C. Shepherd. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4 (3):1383 – 1402, 2010. .

[16] I. W. Renner and D. I. Warton. Equivalence of MAXENT and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281, 2013. .

[17] M. Berman and T. R. Turner. Approximating point process likelihoods with GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):31–38, 1992.

[18] N. Cressie. *Statistics for spatial data.* John Wiley & Sons, New York, 1993.

[19] R. W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5:345, 1962.

[20] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. .

[21] G. M. Harris, G. J. Russell, R. I. van Arde, and S. L. Pimm. Rules of habitat use by elephants *Loxodonta africana* in southern Africa:in sights for regional management. *Oryx*, 41(42 (1)):66–75, 1992.

[22] R. Zidon, S. Garti, W. M. Getz, and D. Saltz. Zebra migration strategies and anthrax in Etosha National Park, Namibia. *Ecosphere*, 8(8):e01925, 2017. .

[23] P. Lindeque and P. C. B. Turnbull. Ecology and epidemiology of anthrax in the Etosha National Park, Namibia. *The Onderstepoort Journal of Veterinary Research*, 61 1:71–83, 1994.