







RESEARCH ARTICLE

Realising the Promise of Large Data and Complex Models

Estimating the abundance of a group-living species using multi-latent spatial models

Colin J. Torney¹  | Megan Laxton¹  | David J. Lloyd-Jones²  | Edward M. Kohi³  |
 Howard L. Frederick⁴  | David C. Moyer⁵ | Chediell Mrisha⁶ | Machoke Mwita³ |
 J. Grant C. Hopcraft⁷ 

¹School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; ²FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Cape Town, South Africa; ³Conservation Information Monitoring Unit, Tanzania Wildlife Research Institute, Arusha, Tanzania; ⁴Tanzania Wildlife Research Institute, Arusha, Tanzania; ⁵Integrated Research Center, The Field Museum of Natural History, Chicago, IL, USA; ⁶Ministry of the Natural Resources and Tourism, Dodoma, Tanzania and ⁷Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, UK

Correspondence

Colin J. Torney

Email: colin.torney@glasgow.ac.uk

Funding information

James S. McDonnell Foundation

Handling Editor: Rachel McCre

Abstract

1. Statistical models use observations of animals to make inferences about the abundance and distribution of species. However, the spatial distribution of animals is a complex function of many factors, including landscape and environmental features, and intra- and interspecific interactions. Modelling approaches often have to make significant simplifying assumptions about these factors, which can result in poor model performance and inaccurate predictions.
2. Here, we explore the implications of complex spatial structure for modelling the abundance of the Serengeti wildebeest, a gregarious migratory species. The social behaviour of wildebeest leads to a highly aggregated distribution, and we examine the consequences of omitting this spatial complexity when modelling species abundance. To account for this distribution, we introduce a multi-latent framework that uses two random fields to capture the clustered distribution of wildebeest.
3. Our results show that simplifying assumptions that are often made in spatial models can dramatically impair performance. However, by allowing for mixtures of spatial models accurate predictions can be made. Furthermore, there can be a non-monotonic relationship between model complexity and model performance; complex, flexible models that rely on unfounded assumptions can potentially make highly inaccurate predictions, whereas simpler more traditional approaches involve fewer assumptions and are less sensitive to these issues.
4. We demonstrate how to develop flexible spatial models that can accommodate the complex processes driving animal distributions. Our findings highlight the

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

importance of robust model checking protocols, and we illustrate how realistic assumptions can be incorporated into models using random fields.

KEYWORDS

aerial surveys, animal abundance, multi-latent fields, spatial modelling

1 | INTRODUCTION

Accurately and efficiently estimating the abundance and distribution of organisms is one of the most fundamental components of ecological research. For many species access to resources is becoming increasingly constrained either because contiguous habitats are becoming fragmented (Fischer & Lindenmayer, 2007; Løvschal et al., 2017), or the overall availability of resources is declining due to human use (Ogutu, 2016; Tilman, 2017). In addition, populations of many commercially harvested species are being exploited to the point of extirpation (Barnosky, 2011), and in some cases sudden outbreaks of disease cause massive declines (Milner-Gulland, 2015; Roelke-Parker et al., 1996). The combination of these pressures raises serious concerns about the long-term viability of many species and underscores the importance of getting rapid, accurate estimates of abundance. In most cases, practitioners and managers want to know not only the abundance of animals but also the drivers of long-term changes (such as the impacts of climate effects, over-exploitation or displacement) and also identify bottlenecks at specific life-history stages. This requires repeated estimates of abundance over time.

Ecological surveys are the primary tool for estimating abundance. Although the design of a survey differs depending on the species (Hedges & O'Brien, 2012), in many cases the abundance is calculated by estimating the number of animals per unit area from a series of sampling points and inferring the density over the entire survey area (Norton-Griffiths, 1973). Samples are generally collected along a systematic pre-selected grid or transects that ensure maximal coverage and appropriate statistical power. Therefore, sampling requires adhering to a strict protocol; however, the practicalities of achieving this in the field can be very challenging (for example, transient obstacles such as low cloud-cover can force breaks in aerial transects and therefore sampling becomes discontinuous). Furthermore, observers may see groups of animals occurring off-transect which cannot be included in the enumeration, even if the observations are important. In many circumstances, a survey is constrained to a single species or a group of species that occur sympatrically resulting in multiple surveys required in areas that are rich in biodiversity.

The difficulties of estimating species abundance are compounded when organisms are highly aggregated. Many species occur in groups either because individuals distribute themselves based on the availability of resources in a patchy environment or because social interactions attract individuals to conspecifics (Krause & Ruxton, 2002; Westley et al., 2018). In addition, animals whose distributions are highly dynamic and subject to rapid fission-fusion processes (such as migratory species like wildebeest) pose a particular problem for

aerial surveys because not only do they form fluid yet highly aggregated groups, the time window for conducting a population estimate is often very brief. This restricts the window of opportunity and limits the number of samples that can be feasibly collected to estimate the population size (i.e. the number of point observation, photos or transect). Furthermore, if the window is missed the survey cannot be repeated until the following year because animals have moved or dispersed, which places pressure on enumerators to be efficient and responsive.

The requirements of statistical models are therefore several-fold. Primarily, statistical methods must accurately infer population abundance from incomplete observations even when spatial distributions are highly heterogeneous and subject to high degrees of autocorrelation. Where possible models must be sufficiently flexible to account for irregular survey design and the realities of the data collection procedure, thereby taking pressure off ecologists and wildlife managers in the field. Models must also be complex enough to make full use of the available data and maximize the accuracy of estimates while minimizing their uncertainty. Finally, models must be practical; fitting models to relatively large datasets should be achievable within reasonable time-scales and ideally implemented within well-supported open source software libraries and packages.

Surveys of animal aggregations often have strong spatial autocorrelation between observations. One approach to reduce the effect of spatial autocorrelation is to use the samples to calculate the mean density per transect or stratum and use this summary to calculate abundance, as in the Jolly II method (Jolly, 1969). In this case, all observations within a transect are assumed to be dependent (i.e. autocorrelated), but transects are assumed to be independent. The sample size is defined by number of transects (not the total number of observations across all transects), which can reduce the statistical power.

Recent developments in this area have focused on the use of latent Gaussian random fields (GRFs; Hristopulos, 2020; Rue & Held, 2005). Random fields are a major component of spatial statistical analysis and can be used to represent spatial and spatiotemporal correlation structures which remain unexplained by environmental variables. In this context, a likelihood is specified for an observed response variable and the random field defines the mean, or other parameter, of the likelihood function. A GRF is defined by a mean and a covariance function and has the property that any finite collection of points follows a multivariate Gaussian distribution with the specified mean and covariance. Hence, the covariance function determines the similarity between two locations of the field and is often assumed to be stationary and isotropic, meaning the similarity

between two locations depends only on their distance, and not their relative direction or absolute position.

Latent Gaussian fields are explicitly or implicitly used in many spatial analysis methods, including INLA (Rue et al., 2009), kriging (Cressie, 1990), Gaussian process (GP) regression (Rasmussen & Williams, 2006), and spatial GAMs with smoothing splines (Wood, 2020). In all these implementations a key challenge is appropriately specifying the smoothness properties of the spatial field in a manner that accurately captures the autocorrelation of the data without overfitting to the noise.

In abundance modelling, latent GRFs have previously been used to account for highly spatially correlated data, as well as irregularly spaced sampling frameworks (Cavieres et al., 2021; Cavieres & Nicolis, 2018; Drexler & Ainsworth, 2013; Ver Hoef & Jansen, 2015). Accounting for spatial structuring due to social behaviour, density dependence, or other effects unexplained by environmental variables has been found to improve certainty of abundance estimation. This is of particular importance in management and conservation of keystone species, and in estimating quotas for stock sustainability in fisheries (Calderon-Aguilera et al., 2019). GRFs have also been shown to be effective in dealing with data containing excess zeros when incorporated with zero-inflated or hurdle models (Welsh et al., 1996). This approach can be implemented using multiple random fields where one field defines the probability of the presence or absence, and the second field determines abundance conditional on the species being present (Barry & Welsh, 2002; Sadykova et al., 2017; Smith et al., 2019).

In this work, we explore a scenario in which spatial models are used to improve the accuracy of abundance estimates for an aggregated species. Our approach uses multiple random fields to estimate the abundance of wildebeest in the Serengeti National Park from aerial survey data. We specifically investigate the importance of model complexity and how it interacts with the spatial patterns that arise due to the ecological and behavioural dynamics of the wildebeest (Torney et al., 2018).

Wildebeest are a keystone species in the Serengeti. Each year, approximately 1.3 million wildebeest move through the range of the Greater Serengeti Ecosystem, a round-trip of >1,500 km (Hopcraft et al., 2015; Thirgood et al., 2004) that covers almost the entire 25,000 km² extent of the ecosystem. During the migration, herds move constantly and vary in size, from a few individuals to hundreds of thousands of animals (Hopcraft et al., 2014). Without the annual migratory cycle, much of the region's biodiversity would likely change (Dobson et al., 2010; Hopcraft et al., 2015), as the passage of the migrants affects virtually every other ecological interaction in the ecosystem from below ground nutrient cycling, to insects and avifauna abundance, to predator-prey interactions, to disease regulation (Sinclair et al., 2015).

Despite their large numbers, wildebeest operate at or close to their physiological limits. Historically, drastic population declines have been observed due to drought, access to forage and disease (Mduma et al., 1999). In adult females, the energetic demands of reproduction are continual and acute as the cycle of pregnancy and

lactation encompasses a full calendar year; calving is highly synchronous and coincides with the period of peak primary production in February (Hopcraft et al., 2014). Approximately 400,000 calves are born each year of which about 50% survive to reproductive age (Hopcraft et al., 2015). The spatial diversity of the ecosystem means that at each point in the migration the quality and abundance of forage changes, as does the exposure to predators, as well as the animals' proximity to human infrastructure and poaching. As many as 100,000 wildebeest are illegally harvested every year (Rentsch & Packer, 2014), and there are upwards of 3,500 tourists per day entering the Serengeti. These anthropogenic inputs create additional hazards for animals to navigate as they strive to gain sufficient energy to survive and reproduce whilst avoiding predation. Any changes to the mass migration are likely to have severe environmental consequences as the wildebeest play a vital role in the ecosystem.

Estimating the abundance of wildebeest in the ecosystem is vital for the effective management of the park and is perhaps the most important metric of the ecosystem's health (Estes, 2014). To estimate the abundance of the wildebeest population, transects are flown over the herds in March, April or May (Campbell & Borner, 1995; Norton-Griffiths, 1973) while the majority of the wildebeest are on the short grass plains in the south-east of Serengeti and the Ngorongoro conservation area, before the migration moves into the woodland areas of the western Serengeti. Transects are flown East to West at 2.5 km intervals with the start and end of each transect determined by the presence of wildebeest. During the transects nadir georeferenced aerial photographs are taken of the survey area at fixed 10-s intervals with the aircraft travelling at close to constant speed at about 1,000 feet above the ground. Once the images are collected the next stage of the process is to identify and count all wildebeest within each image (Torney et al., 2019).

To estimate population abundance from image counts, Jolly's method II (Jolly, 1969) is used. This provides an expected abundance calculated by multiplying the average within image density by the total survey area and a standard error calculated based on the standard deviation of the densities across each transect (the area of each image is calculated using the field of view of the lens and the altitude above-ground). It should be noted that the photographed area is typically less than 3% of the total survey area and typically between 5,000 and 10,000 wildebeest are manually identified out of a population total of around 1.3 million.

Here we compare the Jolly II method used to infer total wildebeest abundance with two hierarchical models that incorporate random spatial fields. We implement two spatial models; the first uses a single field to model the log intensity of a Poisson distribution for wildebeest abundance, while the second introduces a multi-latent model that makes use of a spatial mixture model with multiple random fields to account for the large regions of the survey area that contain zeros in the image counts. To evaluate the performance of these models we generate synthetic wildebeest distributions and vary the fraction of the landscape (denoted p) where wildebeest are present. Through our analysis of inferred wildebeest abundance we show that there is a non-monotonic relationship between model

complexity and accuracy, as large errors are observed using a complex model when model assumptions are not met.

2 | MATERIALS AND METHODS

2.1 | Single-field model

Our first approach to incorporating a spatial field uses a log Cox GP. The distribution of wildebeest is therefore, modelled as an inhomogeneous Poisson process with spatially varying intensity $\lambda(\mathbf{x})$. A GP is used to model the log of the spatially varying intensity so that the full model for wildebeest density D_w is defined as

$$\begin{aligned} D_w(\mathbf{x}): & \text{Poisson}(\lambda(\mathbf{x})) \\ \lambda(\mathbf{x}) & = \exp(F(\mathbf{x})) \\ F(\mathbf{x}): & \text{GP}(\mu, K(\mathbf{x}, \mathbf{x}')), \end{aligned} \quad (1)$$

where GP defines a GP prior, μ is the mean of the log intensity field and $K(\mathbf{x}, \mathbf{x}')$ is a stationary covariance function that determines the similarity of two locations based on their distance.

Several methods are available for parameter inference and model fitting, including MCMC, variational inference and Integrated Nested Laplace Approximation (INLA; Rue et al., 2009). The appropriateness of different approaches will depend on the choice of covariance function and dataset size. Here we assume a Matérn covariance function that allows us to take advantage of the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011). The solution to a given GRF with Matérn covariance is an SPDE, which in this approach is used to estimate a corresponding Gaussian Markov random field (GMRF). The GMRF can then be used to represent the GRF in computation, which due to its Markov property, considerably reduces computational cost, enabling complex spatial and spatiotemporal modelling within an efficient timeframe. Inference is implemented within *inlabru* (Bachl et al., 2019).

2.2 | Multi-latent model

A notable characteristic of the wildebeest distribution is that they are highly aggregated in certain regions of the survey area. Due to the survey design and the constraints of flying transects, this leads to large regions that contain no wildebeest. Counts of wildebeest with aerial photos, therefore, contain large numbers of zeros, which are well known to cause problems when fitting to Poisson distributions (Harrison, 2014).

A standard approach to dealing with zero-inflated data is to use a mixture model that combines a Bernoulli distribution with a Poisson or negative binomial distribution; in this approach the count is zero with a fixed probability, otherwise it is drawn from the discrete probability distribution that may be truncated to exclude zero (Wenger & Freeman, 2008). Although models of this type are better able to cope with zero inflated data, they do not account for the

spatial autocorrelation in the presence of zeros. In the context of the wildebeest distribution, zero counts are highly correlated with one another and there are large clusters of dense and empty regions. It is unlikely these clusters are driven by environmental factors, instead they are driven by the social behaviour of the wildebeest and their aggregation into herds of various sizes.

To account for the spatial clustering of wildebeest we use a multi-latent random field model (Saul et al., 2016). The model consists of two spatial fields, both modelled as latent GPs. The first field determines whether wildebeest are present in the local area or not, whereas the second field is used to model the log intensity of the wildebeest distribution, as in the single-field approach. Hence, the spatial distribution of wildebeest is determined by two independent random fields acting as a mixture as in a hurdle or zero-inflated model; one that determines presence or absence, and one that determines density given wildebeest are present. The model is then defined as follows:

$$\begin{aligned} D_w(\mathbf{x}): & \begin{cases} \text{Poisson}(\lambda(\mathbf{x})), & \text{if } \theta(\mathbf{x}) = 1 \\ 0, & \text{if } \theta(\mathbf{x}) = 0 \end{cases} \\ & \lambda(\mathbf{x}) = \exp(F(\mathbf{x})) \\ \theta(\mathbf{x}): & \text{Binomial}(\text{sigmoid}(G(\mathbf{x}))) \\ F(\mathbf{x}): & \text{GP}(\mu_f, K_f(\mathbf{x}, \mathbf{x}')) \\ G(\mathbf{x}): & \text{GP}(\mu_g, K_g(\mathbf{x}, \mathbf{x}')), \end{aligned} \quad (2)$$

where, as in the single-field model, \mathbf{x} is the spatial location, $D_w(\mathbf{x})$ is the density of wildebeest, which is modelled as an inhomogeneous Poisson process with intensity $\lambda(\mathbf{x})$, and $F(\mathbf{x})$ is a GP that is used to model the log of the spatially varying intensity. $G(\mathbf{x})$ is a second, independent GP that is used to model the binary variable $\theta(\mathbf{x})$ which determines whether wildebeest are present. This variable is drawn from a Binomial distribution where $G(\mathbf{x})$ determines the logit of the presence probability.

Given the additional complexity of the model, computational inference is more challenging. We again use an approximate method, this time using variational inference (Blei et al., 2017) to estimate the posterior distribution of the latent spatial fields. Variational inference follows a similar approach to the Laplace approximation used by INLA; however, it proceeds by firstly assuming a specific form for the posterior density and then minimizing the Kullback–Leibler divergence between the posited density and the target density. Variational inference is a popular method within the machine learning community for fitting GP models (Hensman et al., 2013; Opper & Archambeau, 2009; Titsias, 2009) and has recently been applied to multi-latent GPs such as the model defined by Equation 2 (Saul et al., 2016). We follow this method (Torney, 2022) and use an implementation provided by the package *GPflow* (Matthews et al., 2017).

2.3 | Synthetic data

To test the various inference schemes we generate synthetic wildebeest distributions and then simulate the survey by making

observations of the generated distribution within small areas along a sequence of transects. This process generates a set of counts corresponding to the aerial photographs collected during the real survey.

To generate the synthetic distributions that match the clustering of the real wildebeest, we generate two random fields. The first field is used to define a region where wildebeest are potentially present. We define this field to have a correlation length of 5 km and specify a fraction p of the survey area that may contain wildebeest. This fraction is varied from $p = 0.1$ (i.e. 10% of the area may contain wildebeest) to $p = 1$ (i.e. the entire area may contain wildebeest) in steps of 0.1. For areas that may contain wildebeest, the second field defines the intensity parameter of a Poisson distribution with a mean function specified so that there are approximately 1.3 million wildebeest in total in the region. Random samples are then generated to create a distribution of wildebeest. The process is repeated 10 times for each value of the presence fraction p . See Figure 1 for example distributions and image counts generated for two different values of p .

2.4 | Empirical data

We further assess the multi-latent field approach using survey data from the 2015 wildebeest count. The 2015 count was conducted between 23 April and 2 May over the eastern and southern plains of Serengeti National Park, Ngorongoro Conservation Area, Loliondo

Game Controlled Area, and Maswa Game Reserve. The 2015 aerial survey of the Serengeti National Park and Ngorongoro Conservation Area was undertaken by the Frankfurt Zoological Society (FZS) in conjunction with TAWIRI (Tanzania Wildlife Research Institute).

A Cessna C182 aircraft was used to conduct 10.3hr of photographic sampling flights along east-west transects that covered a straight-line distance of 2,040km. Photographs were taken using a NIKON D800 through a 35-mm Nikor Lens. The camera was mounted in a port in the floor of the aircraft and an automatic trigger was used to collect images at 10-s intervals during each transect. In total 1,584 georeferenced images were taken with a resolution of $7,360 \times 4,912$ pixels, which were later viewed and the number of wildebeest in each image was recorded.

3 | RESULTS

3.1 | Synthetic data

To investigate the accuracy and precision of the different inference procedures we apply the three methods, Jolly II, the single field Poisson likelihood model, and the multi-latent field model, to simulated data. For each dataset, we use each method to calculate the estimate of abundance and the uncertainty associated with the estimate. We then compare the estimate of abundance with the ground truth value.

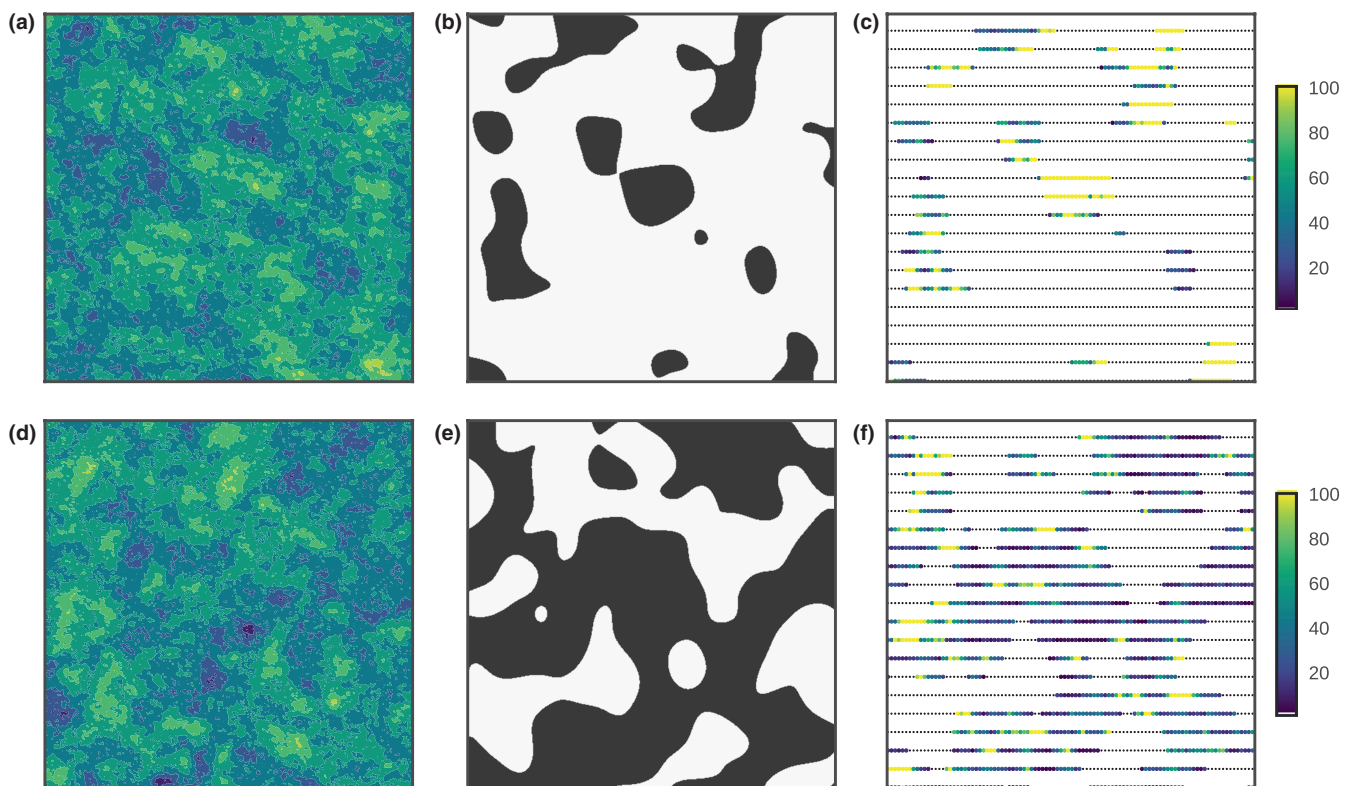


FIGURE 1 Simulated data generation. (a) Generated spatial log-intensity field used to create counts of wildebeest if present. (b) Random presence-absence field for $p = 0.2$. (c) Wildebeest counts generated by flying in silico transects and collecting aerial survey images at 10s intervals. (d–f) as first row with $p = 0.6$.

Figure 2 shows the results from each approach. Examining these results reveals a non-monotonic relationship between model complexity and accuracy. Although both the Jolly II approach and the single latent field assume a homogeneous distribution of wildebeest, the consequences for the more complex model are far more severe. The single field model provides accurate predictions only when model assumptions are met (i.e. $p = 1.0$). When wildebeest are clustered and the survey region contains large numbers of zero image counts, the model overfits the data and results in abundance estimates that are several orders of magnitude greater than the true total. This issue could in principle be overcome through more stringent model regularization techniques, stronger priors on the hyperparameters, or preprocessing of the survey data; however, it is notable that the traditional method of Jolly II provides a robust and reliable estimate.

The multi-latent approach displays a similar accuracy to the Jolly II method; however, it is less biased than Jolly II, which systematically appears to result in an undercount, and it is better able to quantify the uncertainty in the abundance. For all values of the fraction of the landscape potentially occupied by wildebeest, the true value was more likely to fall within the 95% credible intervals when compared to the 95% confidence interval of the Jolly method. The model was also able to accurately capture the latent presence-absence field used to generate the data as shown in Figure S4. It is likely that the uncertainty quantification of the multi-latent model could be further improved by careful selection of the covariance function used to encode spatial autocorrelation in each field. For the simulation study a Matérn 3/2 kernel was used for both fields but other functions that better approximate observed spatial patterns would likely result in greater accuracy.

3.2 | Empirical data

We next apply the multi-latent model to the 2015 wildebeest image counts. The spatial distribution of image counts is shown in Figure 3a. In this year, wildebeest were distributed across the landscape into several clusters with large regions of zero counts. After optimizing the model, the learned spatial fields define the probability of wildebeest being present, and their log-density given wildebeest are present. Figure 3b shows the presence probability for the 2015 data, whereas Figure 3c displays the log-intensity.

The optimized model parameters for the lengthscale of the presence-absence field and the log-intensity field are shown in Table 1 along with the population estimates for the year obtained using the multi-latent model and, for comparison, Jolly II. The posterior distribution of total wildebeest abundance is shown in Figure 4 with lines indicating the posterior median and the upper and lower 95% credible intervals. Again, the Jolly II outputs are shown for comparison. Based on the results from the simulation study, we expect that the true abundance is underestimated by Jolly II and the credible intervals obtained from the multi-latent model are a more accurate reflection of the uncertainty in the estimates.

4 | DISCUSSION

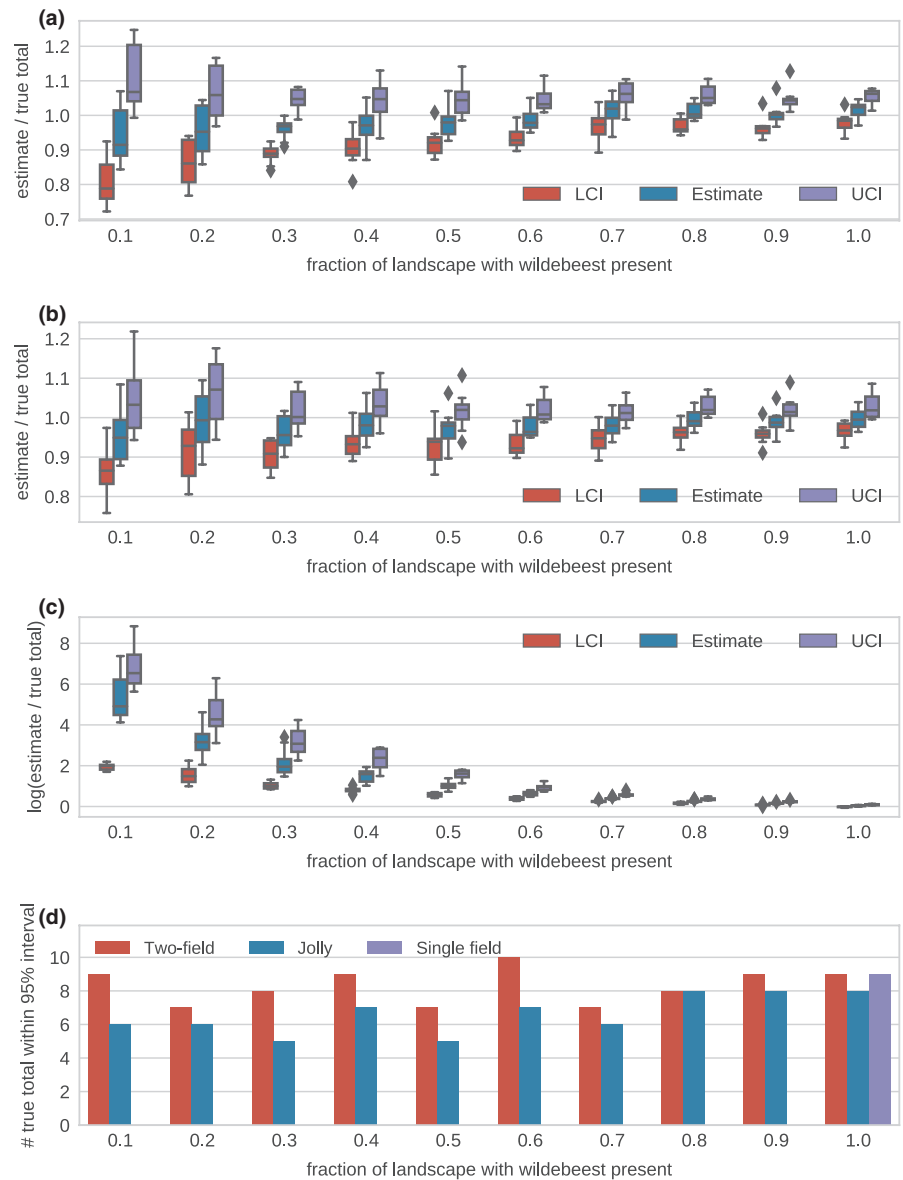
Our results show that an increase in model complexity does not necessarily lead to improved model performance when the assumptions of the analysis are violated. Often the simplest models (such as Jolly II) rely on fewer assumptions and are less flexible but are robust, as long as data are collected systematically and methodically. The inclusion of more complexity into models such as the single field model in which the observations can be spatially autocorrelated introduces assumptions about stationarity of the covariance structure. When these assumptions were violated, the single-field model performed poorly in comparison to the more traditional Jolly II approach; this was especially pronounced under conditions of strong clustering of wildebeest herds where abundance estimates were severely over-inflated. However, the inclusion of further complexity in the multi-latent model reconciles this issue by allowing two random fields that determine presence-absence separately from density (i.e. a hierarchical approach to spatial autocorrelation in which concentrated aggregations can occur within groups that are distinct from each other). This results in a comparable level of accuracy in abundance estimates with the Jolly II method. Furthermore, the multi-latent model suffers from less bias than Jolly II and provides a more accurate quantification of uncertainty. Therefore, the inclusion of complexity in a model does not necessarily equate to improved performance, but increasingly complex models that make reasonable assumptions can provide advantages over simpler, more traditional approaches.

An extension to multi-latent field framework is to incorporate the effects of environmental features (see Figure S1). The use of environmental information allows the model to infer the relationship between known covariates and the species distribution and then use this relationship to make more informed predictions about density in unsampled regions. This approach is straightforward to implement via a modification to the GRF but could not be introduced into Jolly II due to its use of transect as the unit of observation.

The objective of species abundance models is to maximize both accuracy and flexibility while minimizing the computational time, which allows for fast, repeatable and precise measures of distribution and abundance. There is an inherent trade-off within these aims, which means the choice of model should depend on the scale of improvement and the advances it makes in terms of conservation impact. Here, we discuss three advances offered by increased model complexity when estimating abundance that would be relevant to other group living or migratory animals.

Although the results of the most complex (multi-latent) model and the simplest approach (Jolly II) were comparable, increased model complexity can allow for practical on-the-ground advantages. For instance, Jolly II constrains enumerators to fly transects along straight lines with systematic sampling (i.e. photographs are taken every 10 s). The multi-latent approach allows for greater flexibility in sampling, as it uses the coordinate locations of photographs, rather than overall transects,

FIGURE 2 Comparison of methods on simulated data. (a) Population abundance estimate and 95% upper and lower credible interval using the multi-latent field approach. Results show model accuracy for 10 independent simulated datasets with wildebeest occupying various fractions of the synthetic survey area. Estimates are rescaled to be a fraction of the true population size so that a value of 1.0 means the estimate matches the true abundance. (b) Population abundance estimate for the same simulated data using Jolly's method II. (c) Population estimate using a single field when the data is generated with a multi-latent field. Note the accuracy is now showing the log of the ratio between the true and estimated value. Diamond markers show outliers that lie beyond 1.5 times the interquartile range. (d) Number of simulation runs where the true value was between the model's 95% uncertainty interval. For the single field approach the true value was outside of the credible interval in all simulations except when $p = 1.0$.



to account for observation effort. Using this approach to estimate abundance means that enumerators are not constrained to transects or systematic sampling. We show how the multi-latent model is unaffected by changes in the sampling strategy by performing random sampling of the area and a spiral survey path in place of the traditional transect pattern (see [Figures S2](#) and [S3](#)). This provides flexibility for enumerators to tackle obstacles in data collection such as navigating the aircraft around transient barriers (e.g. thunder-heads). If the sampling does not strictly follow transects, then enumerators need to pay attention to the GPS tracklog to avoid resampling areas and double counting animals.

Greater sampling flexibility would also allow enumerators to respond to the current distribution of animals encountered during a survey. This enables a more stratified sampling approach in which effort can be focused based on the density of animals (e.g. greater sampling effort in high-density areas, and lower in low-density areas). Concentrating the sampling in this way could potentially

improve abundance estimation, as more information is required to extract accurate counts in high-density areas, whereas relatively little information (i.e. fewer photographs at larger intervals) is needed to determine the absence of the species. This flexibility could create a more efficient approach to data collection, allowing enumerators to make sampling decisions in response to the current distribution of animals.

Flexible sampling could also allow for more efficient data collection by enabling multiple species to be counted in a single survey. In the traditional approach using Jolly II, transects are predetermined according to the distribution of the wildebeest herds. However, zebra *Equus quagga*, Thomson's gazelle *Eudorcas thomsonii*, Grant's gazelle *Nanger granti* and eland *Taurotragus oryx* also migrate to the same area. Herds of different species tend to overlap but their distributions are not identical, thus traditionally enumerators could only count wildebeest during the survey (separate surveys focused on other species). However, the greater sampling flexibility introduced by the multi-latent approach is geographically unconstrained and,

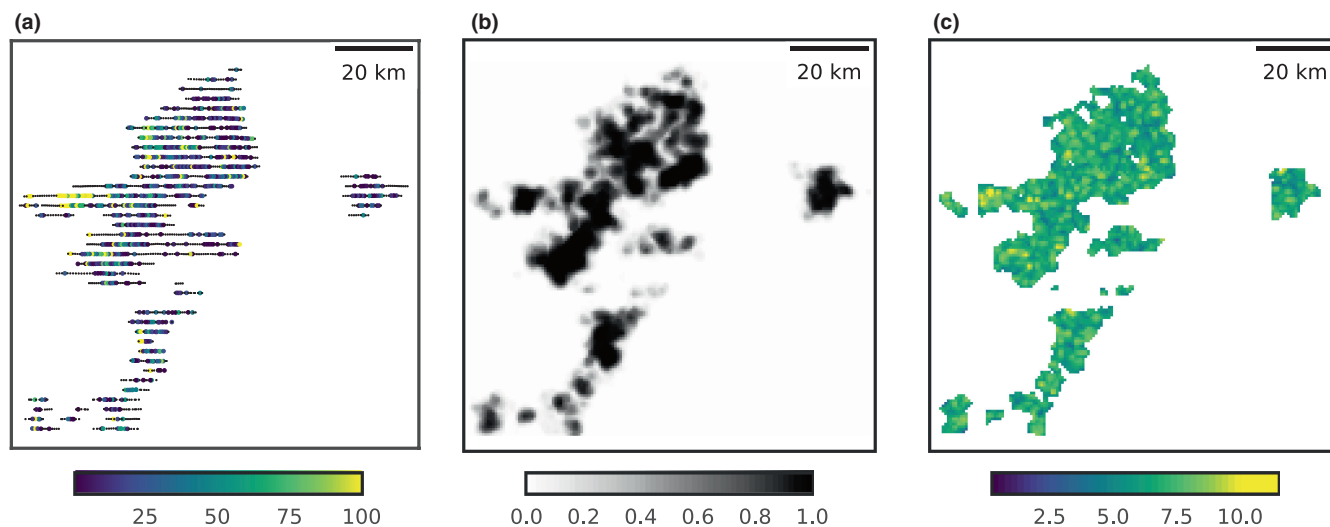


FIGURE 3 Inference of wildebeest abundance. (a) Wildebeest counts from images collected as part of the 2015 survey. Colours indicate the count in each image; photo locations without wildebeest present are shown as black points. (b) Inferred presence-absence field. Values close to 1 indicate a high-probability wildebeest are present in the area. White regions indicate wildebeest are absent. (c) Log-intensity field. Values shown indicate the expected log density of wildebeest. Areas where wildebeest are absent with probability >0.9 (as determined by (b)) are omitted.

TABLE 1 Inferred parameters and model outputs for the 2015 survey data

Parameter/output	Value
Presence field lengthscale	3.79 km
Log-intensity lengthscale	0.46 km
Posterior mean abundance	1,351,045
Posterior median abundance	1,343,906
Lower 95% CI	1,177,577
Upper 95% CI	1,548,680
Jolly II est. abundance	1,326,709
Jolly II Lower 95% CI	1,152,930
Jolly II Upper 95% CI	1,500,487

therefore, enables sampling over a wider area. There is no need to predetermine sampling transects according to a single species' distribution, enabling the collection of data on multiple species in a single survey. In combination with deep-learning approaches to object recognition in images (Torney et al., 2019), this could be an extremely powerful approach, and potentially represents a significant advance in improving the efficiency of resource-use in counting animals for abundance estimation.

Our results demonstrate the traditional method for inferring wildebeest abundance, Jolly II, is a reliable approach that is relatively straightforward to implement. Given a fixed survey design that adheres to the requirements of the Jolly II method there is no clear advantage to implementing the more complex multi-latent field approach. However, the more complex model is both flexible and extensible. Potential extensions include the use of environmental covariates as described above, as well as explicitly modelling the temporal dynamics of the wildebeest herds (Blangiardo

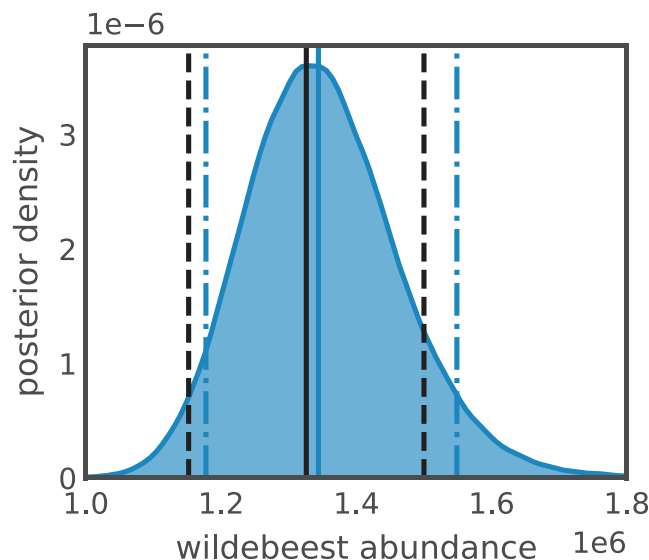


FIGURE 4 Posterior distribution of wildebeest abundance. 100,000 samples from the fitted multi-latent model were taken and the posterior distribution was plotted using a kernel density estimator. Blue lines show posterior median abundance (solid line) and 95% credible interval (broken-dashed line). Black lines show the mean and standard error of the Jolly II estimate.

et al., 2013). The greater flexibility of the model translates to greater flexibility in the field when conducting the survey, as well as the prospect of including multiple species in the analysis and incorporating data from multiple sources (Graves et al., 2022; Robinson et al., 2021).

AUTHORS CONTRIBUTIONS

Colin J. Torney, J. Grant C. Hopcraft, Megan Laxton designed the study; David J. Lloyd-Jones, David C. Moyer, Edward M. Kofi,

Howard L. Frederick, Chediel Mrisha, Machoke Mwita, J. Grant C. Hopcraft collected data; Colin J. Torney, Megan Laxton, J. Grant C. Hopcraft, David J. Lloyd-Jones analysed the data; Colin J. Torney, J. Grant C. Hopcraft, Megan Laxton wrote the manuscript with input from all authors.

ACKNOWLEDGEMENTS

J.G.C.H. acknowledges support from the British Ecological Society large grant scheme, the Friedkin Foundation and the European Union Horizon 2020 grant no 641918. C.J.T. acknowledges support from a James S. McDonnell Foundation Studying Complex Systems Scholar Award. M.L. is supported by an EPSRC scholarship. We are grateful to two reviewers whose helpful comments and suggestions improved the manuscript.

CONFLICT OF INTEREST

We declare that we have no competing interests.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13941>.

DATA AVAILABILITY STATEMENT

Raw count data and source code are available from <https://doi.org/10.5281/zenodo.6684421> (Torney, 2022).

ORCID

Colin J. Torney  <https://orcid.org/0000-0003-1673-7835>

Megan Laxton  <https://orcid.org/0000-0002-5017-7395>

David J. Lloyd-Jones  <https://orcid.org/0000-0001-7880-5659>

Edward M. Kohi  <https://orcid.org/0000-0002-4815-7600>

Howard L. Frederick  <https://orcid.org/0000-0003-0565-0769>

J. Grant C. Hopcraft  <https://orcid.org/0000-0002-9617-9951>

REFERENCES

- Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). Inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6), 760–766.
- Barnosky, A. (2011). Has the earth's sixth mass extinction already arrived? *Nature*, 471, 51–57.
- Barry, S. C., & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2–3), 179–188.
- Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-Temporal Epidemiology*, 4, 33–49.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.
- Calderon-Aguilera, L. E., Graham, L. J., & Eigenbrod, F. (2019). Ignoring the spatial structure of the sea cucumber *isostichopus fuscus* distribution when granting quotas can be costly. *Ocean & Coastal Management*, 178, 104859. <https://doi.org/10.1016/j.ocecoaman.2019.104859>
- Campbell, K., & Borner, M. (1995). Population trends and distribution of Serengeti herbivores: Implications for management. *Serengeti II: Dynamics, Management, and Conservation of an Ecosystem*, 2, 117–145.
- Cavieres, J., Monnahan, C. C., & Vehtari, A. (2021). Accounting for spatial dependence improves relative abundance estimates in a benthic marine species structured as a metapopulation. *Fisheries Research*, 240, 105960. <https://doi.org/10.1016/j.fishres.2021.105960>
- Cavieres, J., & Nicolis, O. (2018). Using a spatiotemporal bayesian approach to estimate the relative abundance index of yellow squat lobster (*Cervimunida johni*) off Chile. *Fisheries Research*, 208, 97–104.
- Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22(3), 239–252.
- Dobson, A. P., Borner, M., Sinclair, A. R. E., Hudson, P. J., Anderson, T. M., Bigurube, G., Davenport, T. B. B., Deutsch, J., Durant, S. M., Estes, R. D., Estes, A. B., Fryxell, J., Foley, C., Gadd, M. E., Haydon, D., Holdo, R., Holt, R. D., Hopcraft, J. G. C., Hilborn, R., ... Wolanski, E. (2010). Road will ruin Serengeti. *Nature*, 467(7313), 272–273.
- Drexler, M., & Ainsworth, C. H. (2013). Generalized additive models used to predict species abundance in the gulf of Mexico: An ecosystem modeling tool. *PLoS ONE*, 8(5), e64458.
- Estes, R. (2014). *The gnu's world: Serengeti wildebeest ecology and life history*. University of California Press.
- Fischer, J., & Lindenmayer, D. B. (2007). Landscape modification and habitat fragmentation: A synthesis. *Global Ecology and Biogeography*, 16(3), 265–280.
- Graves, T. A., Yarnall, M. J., Johnston, A. N., Preston, T. M., Chong, G. W., Cole, E. K., Janousek, W. M., & Cross, P. C. (2022). Eyes on the herd: Quantifying ungulate density from satellite, unmanned aerial systems, and GPS collar data. *Ecological Applications*, e2600–e2616.
- Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2, e616.
- Hedges, S., & O'Brien, T. (2012). Aerial survey methods. In *Monitoring elephant populations and assessing threats* (pp. 162–171). Universities Press (India).
- Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence* (pp. 282–290). AUAI Press.
- Hopcraft, J. G. C., Holdo, R., Mwangomo, E., Mduma, S., Thirgood, S., Borner, M., Fryxell, J., Olff, H., & Sinclair, A. (2015). Why are wildebeest the most abundant herbivore in the Serengeti ecosystem. In *Serengeti IV* (Vol. 125). University of Chicago Press.
- Hopcraft, J. G. C., Morales, J. M., Beyer, H., Borner, M., Mwangomo, E., Sinclair, A. R. E., Olff, H., & Haydon, D. T. (2014). Competition, predation, and migration: Individual choice patterns of Serengeti migrants captured by hierarchical models. *Ecological Monographs*, 84(3), 355–372.
- Hristopulos, D. T. (2020). *Random fields for spatial data modeling: A primer for scientists and engineers*. Springer Nature.
- Jolly, G. (1969). Sampling methods for aerial censuses of wildlife populations. *East African Agricultural and Forestry Journal*, 34(sup1), 46–49.
- Krause, J., & Ruxton, G. D. (2002). *Living in groups*. Oxford University Press.
- Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4), 423–498.
- Løvschal, M., Bøcher, P. K., Pilgaard, J., Amoke, I., Odingo, A., Thuo, A., & Svenning, J.-C. (2017). Fencing bodes a rapid collapse of the unique greater mara ecosystem. *Scientific Reports*, 7(1), 1–7.
- Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrà, P., Ghahramani, Z., & Hensman, J. (2017). Gpflow: A Gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18(40), 1–6.
- Mduma, S. A., Sinclair, A., & Hilborn, R. (1999). Food regulates the Serengeti wildebeest: A 40-year record. *Journal of Animal Ecology*, 68(6), 1101–1122.

- Milner-Gulland, E. (2015). Catastrophe and hope for the saiga. *Oryx*, 49(4), 577.
- Norton-Griffiths, M. (1973). Counting the serengeti migratory wildebeest using two-stage sampling. *African Journal of Ecology*, 11(2), 135–149.
- Ogutu, J. (2016). Extreme wildlife declines and concurrent increase in livestock numbers in Kenya: What are the causes? *PLoS ONE*, 11, e0163249.
- Opper, M., & Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, 21(3), 786–792.
- Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.
- Rentsch, D., & Packer, C. (2014). The effect of bushmeat consumption on migratory wildlife in the Serengeti ecosystem, Tanzania. *Oryx*, 49, 287–294.
- Robinson, C., Ortiz, A., Hughey, L., Stabach, J. A., & Ferres, J. M. L. (2021). Detecting cattle and elk in the wild from space. *arXiv preprint arXiv:2106.15448*.
- Roelke-Parker, M. E., Munson, L., Packer, C., Kock, R., Cleaveland, S., Carpenter, M., O'Brien, S. J., Pospischil, A., Hofmann-Lehmann, R., Lutz, H., Mwamengele, G. L. M., Mgasa, M. N., Machange, G. A., Summers, B. A., & Appel, M. J. G. (1996). A canine distemper virus epidemic in Serengeti lions (*Panthera leo*). *Nature*, 379(6564), 441–445.
- Rue, H., & Held, L. (2005). *Gaussian markov random fields: Theory and applications*. CRC Press.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392.
- Sadykova, D., Scott, B. E., De Dominicis, M., Wakelin, S. L., Sadykov, A., & Wolf, J. (2017). Bayesian joint models with inla exploring marine mobile predator–prey and competitor species habitat overlap. *Ecology and Evolution*, 7(14), 5212–5226.
- Saul, A. D., Hensman, J., Vehtari, A., & Lawrence, N. D. (2016). Chained gaussian processes. In A. Gretton & C. Robert (Eds.), *Proceedings of the nineteenth international workshop on artificial intelligence and statistics* (Vol. 51, pp. 1431–1440). PMLR.
- Sinclair, A. R., Metzger, K. L., Mduma, S. A., & Fryxell, J. M. (2015). *Serengeti IV: Sustaining biodiversity in a coupled human-natural system*. University of Chicago Press.
- Smith, A., Hofner, B., Lamb, J. S., Osenkowski, J., Allison, T., Sadoti, G., McWilliams, S. R., & Paton, P. (2019). Modeling spatiotemporal abundance of mobile wildlife in highly variable environments using boosted GAMLSS hurdle models. *Ecology and Evolution*, 9(5), 2346–2364.
- Thirgood, S., Mosser, A., Tham, S., Hopcraft, J. G. H., Mwangomo, E., Mlengeya, T., Kilewo, M., Fryxell, J., Sinclair, A., & Borner, M. (2004). Can parks protect migratory ungulates? The case of the Serengeti wildebeest. *Animal Conservation*, 7, 113–120.
- Tilman, D. (2017). Future threats to biodiversity and pathways to their prevention. *Nature*, 546, 73–81.
- Titsias, M. (2009, 16–18 Apr). Variational learning of inducing variables in sparse Gaussian processes. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 5, 567–574.
- Torney, C. J. (2022, June). ctorney/multi-latent-fields: Publication release. *Zenodo*. <https://doi.org/10.5281/zenodo.6684421>
- Torney, C. J., Hopcraft, J. G. C., Morrison, T. A., Couzin, I. D., & Levin, S. A. (2018). From single steps to mass migration: The problem of scale in the movement ecology of the serengeti wildebeest. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746), 20170012.
- Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., Kohi, E. M., & Hopcraft, J. G. C. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10, 779–787.
- Ver Hoef, J. M., & Jansen, J. K. (2015). Estimating abundance from counts in large data sets of irregularly spaced plots using spatial basis functions. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1), 1–27.
- Welsh, A. H., Cunningham, R. B., Donnelly, C., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, 88(1–3), 297–308.
- Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10), 2953–2959.
- Westley, P. A., Berdahl, A. M., Torney, C. J., & Biro, D. (2018). *Collective movement in ecology: From emerging technologies to conservation and management* (Vol. 373). The Royal Society.
- Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. *TEST*, 29(2), 307–339.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Torney, C. J., Laxton, M., Lloyd-Jones, D. J., Kohi, E. M., Frederick, H. L., Moyer, D. C., Mrisha, C., Mwita, M., & Hopcraft, J. G. C. (2023). Estimating the abundance of a group-living species using multi-latent spatial models. *Methods in Ecology and Evolution*, 14, 77–86. <https://doi.org/10.1111/2041-210X.13941>