Constructing wildebeest density distributions by spatio-temporal smoothing of ordinal categorical data using GAMs

Elaine A. Ferguson[1], Jason Matthiopoulos[1], Dirk Husmeier[2]

[1] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[2] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, UK

E-mail for correspondence: e.ferguson.2@research.gla.ac.uk

Abstract: Spatio-temporal smoothing of large ecological datasets describing species distributions can be made challenging by high computational costs and deficiencies in the available data. We present an application of a GAM-based smoothing method to a large ordinal categorical dataset on the distribution of wildebeest in the Serengeti ecosystem.

Keywords: GAMs; Ordinal categorical; Smoothing; Spatio-temporal; Wildebeest.

1 Introduction

Spatio-temporal smoothing of species distribution data has many potential uses in ecology; for example, to provide a smooth density function that can be used with gradient matching approaches (Xun et al. 2013) to fit partial differential equation (PDE) models of animal movement. A range of smoothing methods (kernel density estimation, splines, Gaussian processes, etc.) have been developed in the statistical literature. However, the practicalities and expense involved in collecting species distribution data over large areas in the field can mean that the data are not in a form that these methods can readily be applied to. Ordinal categorical data, for example, may be collected when it is infeasible to accurately count all individuals in a population, so that the abundance at each point in space and time is instead estimated as belonging to a broader abundance category. A

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

relatively small number of approaches have been developed for smoothing data of this type, where we need to recover the underlying true density of individuals from the categories (Chu and Ghahramani 2005, Wood et al. 2016). Smoothing large datasets in multiple dimensions can also be made challenging by high computational costs. Methods that allow smoothing of these datasets even when computational resources are limited would therefore be very useful. Here we present an application of a method for applying spatio-temporal smoothing to a large ordinal categorical dataset on the distribution of wildebeest in the Serengeti ecosystem of Tanzania and Kenya.

2 Methods

The wildebeest distribution data, which have been described and utilised in a number of previous studies (Norton-Griffiths 1973, Maddock 1979, Boone et al. 2006, Holdo et al. 2009), were obtained from monthly aerial surveys of the Serengeti ecosystem during the period from August 1969 to August 1972. Each cell in a grid of $25km^2$ cells was assigned to one of five wildebeest abundance categories: 0, 1-25, 26-250, 251-2,500 and >2,500 individuals per $25km^2$. There were 2,576 cells making up the spatial grid, all of which were sampled on 33 occasions during the time period, resulting in a large dataset with a total 85,008 data points.
To smooth the data in time, t, and the two spatial dimensions (x, y) we fitted GAMs (generalised additive models) with a tensor product (composed of cubic regression spline smooths, where overfitting was prevented by penalisation of the integral of the squared second derivatives) between these three variables using the mgcv package (Wood 2011) in R (R Core Team 2015). We used the ordinal categorical GAM method described in Wood et al. (2016), where the linear predictor gives the value of a latent variable, here representing the wildebeest density underlying the ordinal categories. The cut-off points that demarcate the five ordinal categories were specified, and the probability that a point in space and time belongs to a given category equals the probability that the latent variable lies between the corresponding category cut-offs at that point.
In Wood et al. (2016), the latent function can range from $-\infty$ to ∞, but we know that wildebeest density has a minimum 0 and a finite maximum $W_{\max}$. We can introduce these constraints by applying a sigmoidal transformation to the latent function L after the GAM has been fitted, giving a preliminary wildebeest density $\hat{W}$ as follows:

$$\hat{W}(x, y, t) = \frac{W_{\max}}{1 + \exp(-L(x, y, t))} \tag{1}$$

Note that this also required that an inverse sigmoid transform be applied to the category cut-offs $\mathbf{c}$ prior to the GAM fitting:

$$\bar{\mathbf{c}} = -\log\left(\frac{W_{\max}}{\mathbf{c}} - 1\right) \tag{2}$$

$W_{\max}$ was estimated by first assuming that the wildebeest densities in the grid cells assigned to the lower four ordinal categories, which had known upper and lower bounds, were equal to the mid-points of those categories. The sum of the densities in these lower category cells for each month was then subtracted from the total number of wildebeest W_T known to be in the region from a population count in 1971 (Norton-Griffiths 1973). The remaining wildebeest for each month were assumed to be divided evenly between the cells in the highest ordinal category (which was unbounded above) for that month. We took $W_{\max}$ to be the largest wildebeest density estimated for cells in the highest abundance category over all months.
Even after applying sensible upper and lower bounds to the latent function, large fluctuations in the area under $\hat{W}$ (which represents the total number of wildebeest in the region) can occur over time. This is undesirable, since we expect wildebeest numbers to remain relatively stable at W_T over the time period of interest. We therefore consider the normalised wildebeest density $\bar{W}$, where the total number of animals is maintained at W_T by normalising $\hat{W}$ as follows:

$$\bar{W}(x, y, t) = \frac{\hat{W}(x, y, t)\, W_T}{\int \hat{W}(x, y, t) dx dy} \tag{3}$$

Due to computational time and memory constraints, a sufficiently flexible GAM could not be fitted to the entire large dataset simultaneously. We therefore divided the time series into three contiguous intervals and fitted a GAM in (x, y, t) to each interval separately. Each GAM had 20 knots in the marginal smooth in each spatial dimension, and a number of knots in the marginal smooth in time that was equal to the number of time points present in the data subset to which the GAM was fitted (11 or 12). This resulted in the effective degrees of freedom, which are determined by the degree of penalization (selected during fitting) applied to the integral of the squared second derivatives, being considerably lower than the maximum number available, suggesting that the number of knots was sufficient (Wood 2006). The three GAMs were joined together by averaging at the link times l_i $(i \in 1, 2)$, with smoothness being maintained by allowing the influence of each GAM on the others to decline smoothly, according to the parameter σ, as distance from the point of joining increased. For a given point $(\bar{x}, \bar{y}, \bar{t})$, therefore, we obtain a final estimate of wildebeest density W by

$$W(\bar{x}, \bar{y}, \bar{t}) = \bar{W}_{GAM_j}(\bar{x}, \bar{y}, \bar{t}) + \sum_{i=1}^{2} a_i \exp\left(\frac{-(\bar{t} - l_i)^2}{2\sigma^2}\right) m_i(\bar{t}) \tag{4}$$

Here $\bar{W}_{GAM_j}$ is the normalised wildebeest density obtained from the GAM fitted to time interval j, where

$$j = \begin{cases} 1 & \text{if} \quad \bar{t} \leq l_1 \\ 2 & \text{if} \quad l_1 < \bar{t} \leq l_2 \\ 3 & \text{if} \quad \bar{t} > l_2 \end{cases} \tag{5}$$

The a_i are given by

$$a_i\left(\bar{x}, \bar{y}, l_i\right) = \frac{\bar{W}_{GAM_i}\left(\bar{x}, \bar{y}, l_i\right) - \bar{W}_{GAM_{i+1}}\left(\bar{x}, \bar{y}, l_i\right)}{2} \tag{6}$$

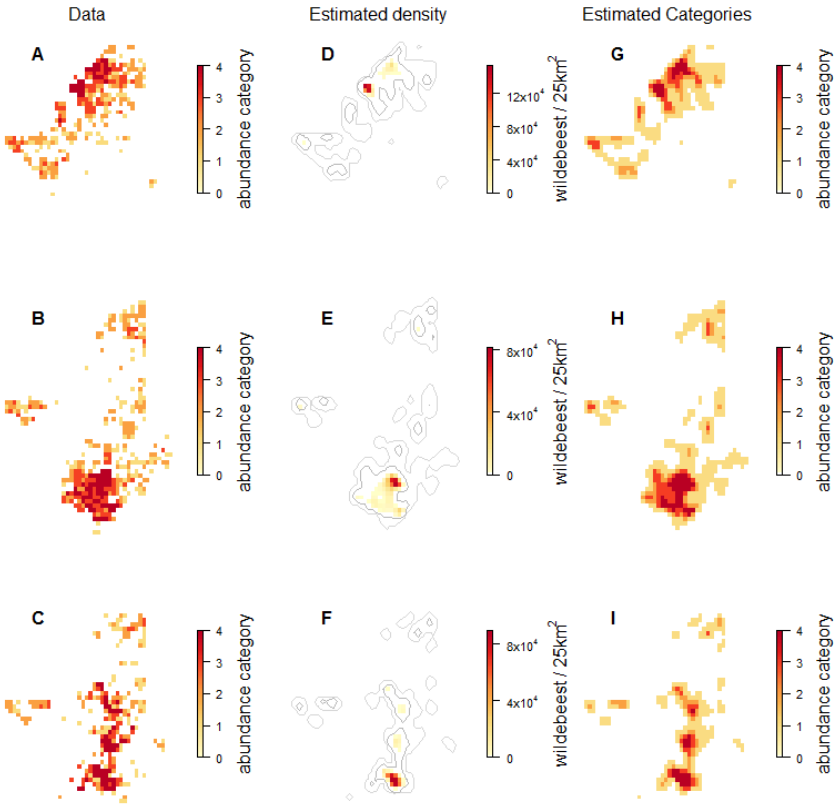

FIGURE 1. Model fit in space at three different time points. **A-C:** The wildebeest spatial distribution data for months 1, 18 and 35. **D-F:** The smooth wildebeest density distribution estimated in space by the model for months 1, 18 and 35. The two contours indicate the boundaries between abundance categories 0, 1 and 2. **G-I:** Estimated wildebeest abundance categories based on **D-F**.

and the m_i, which ensure that the adjustments are made in the correct direction on either side of each link point, are

$$m_i(t) = \begin{cases} -1 & \text{if} \quad \bar{t} \leq l_i \\ 1 & \text{if} \quad \bar{t} > l_i \end{cases} \tag{7}$$

If the influence of the adjoining GAMs declines too slowly with distance from the link points, relative to the rate at which changes occur in $\bar{W}_{GAM_i}$ (i.e. σ is too large), unrealistic negative values of W can occur. We therefore tuned σ by starting with a relatively large value and gradually decreasing it until no negative values of W occurred.

3 Results and Conclusion

The method described was found to successfully produce a smooth function in space that resembles the original data (Figure 1). The resulting function is also observed to be smooth in time, with no evidence that the wildebeest density changes either more slowly or more rapidly around the GAM link times than it does elsewhere in the time period (Figure 2). This suggests that our approach of linking models that have been fitted to subsets of a larger dataset is a promising means of reducing the high computational costs of smoothing large datasets in multiple dimensions. Using this method, we have recovered realistically bounded wildebeest abundance estimates from coarse ordinal categories; an ability that could be useful in the field of ecology where such imperfect data are common. By producing a smooth surface from which spatial and temporal gradients in density can

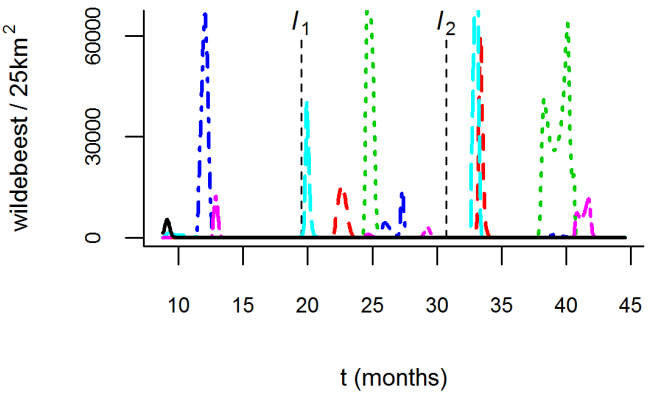

FIGURE 2. Changes in the estimated wildebeest density in six grid cells (indicated by different colours/line types) over the time period of interest. The link times between the three GAMs are indicated by dashed vertical lines.

be calculated, our method also promises to enable statistical inference for PDE models of animal movement using the gradient matching approach of Xun et al. (2013), which we will investigate in future work.

Acknowledgments: Special thanks to Ricardo M. Holdo for providing access to the wildebeest distribution data. E.A.F. is funded by a University of Glasgow Lord Kelvin/Adam Smith PhD scholarship.

References

Boone, R.B., Thirgood, S.J. and Hopcraft J.G.C. (2006). Serengeti wildebeest migratory patterns modeled from rainfall and new vegetation growth. *Ecology*, **87**, 1987 – 1994.

Chu, W. and, Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, **6**, 1019 – 1041.

Holdo, R. M., Holt, R. D. and Fryxell, J. M. (2009). Opposing rainfall and plant nutritional gradients best explain the wildebeest migration in the Serengeti. *The American Naturalist*, **173**, 431 – 445.

Maddock, L. (1979). The migration and grazing succession. In: *Serengeti: dynamics of an ecosystem*, Sinclair, A. R. E., and Norton-Griffiths, M. (editors), Chicago: University of Chicago Press, 104 – 129.

Norton-Griffiths, M. (1973). Counting the Serengeti migratory wildebeest using two-stage sampling. *East African Wildlife Journal*, **11**, 135 – 149.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73**, 3 – 36.

Wood, S.N., Pya, N., and Sfken, B. (2016). Smoothing parameter and model selection for general smooth models. arXiv:1511.03864v2.

Xun, X., Cao, J., Mallick, B., Maity, A., and Carroll, R.J. (2013). Parameter Estimation of Partial Differential Equation Models. *Journal of the American Statistical Association*, **108**, 1009 – 1020.