Bayesian Gaussian Process Beta Regression models for the social sciences: a case study on spatial proportion data from archaeology

Jasmine Vieri [1] Enrico R. Crema[1] Marcos Martinón-Torres [1]

[1]McDonald Institute for Archaeological Research, University of Cambridge

Spatial proportion data

The social sciences often analyse data on rates and proportions influenced by processes that operate spatially — for instance, the graduation rates of high schools within a region, or data on the chemical compositions of archaeological artefacts from different recovery locations. Spatial proportion data, however, violate several key assumptions inherent to simple linear regression:

- **Normality and homoskedasticity**: Proportion data are bounded so that their constituents sum up to a constant, often resulting in skewness and heteroskedasticity.
- **Sample independence**: "Everything is related to everything else, but near things are more related than distant things" (***The First Law of Geography***, Tobler, 1970)

To alleviate issues to do with skewness or asymmetry, proportion data are frequently log-transformed and modelled using the normal distribution. While the approach *can* alleviate skewness in the data, it fails to properly account for the inconsistent variances of proportional response variables.

Similarly, while multilevel modelling is frequently adopted to account for sample interdependence, the approach treats spatial units as nominal categories, therefore not considering spatial autocorrelation structures. To model sample covariance across space, ICAR and Gaussian processes have sometimes been adopted by researchers.

Here, we put forward models that properly account for ***both*** the constrained nature of proportion data ***and*** explicitly allow for the modelling of spatial autocorrelation structures.

Beta Regression

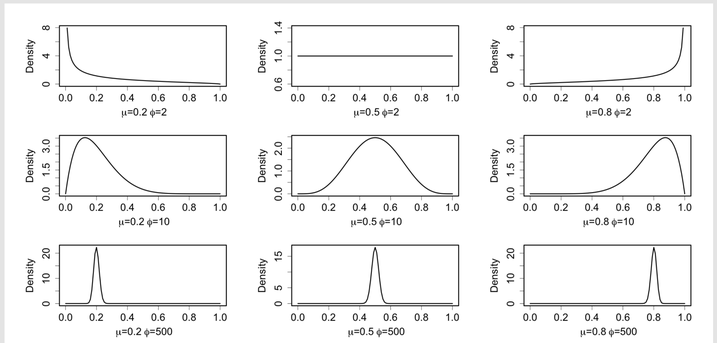

Figure 1. Beta densities at different values of μ and ϕ

Adopting the beta distribution to model proportional responses provides a statistically more robust solution compared to data transformations, for the following reasons:

a) Ensures that no predictions are made outside of the interval (0,1).
b) Accommodates for skewed, U-shaped, uniform and quasi-normal distributions for the outcome.
c) Allows for natural heteroskedasticity in the model.
d) Does not require data transformations, whereby the model outputs are readily interpretable at the original compositional scale (Ferrari & Cribari-Neto, 2004).

Gaussian Process Beta Regression

$$y_i \sim Beta(\mu_i, \phi)$$

$$\text{logit}(\mu_i) = X_i \times \beta + k_{\text{cluster}(i)}$$

$$\begin{pmatrix} k_1 \\ k_2 \\ \dots \\ k_{cluster} \end{pmatrix} \sim GP\left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, K\right)$$

$$K_{ij} = \eta^2 exp - \frac{|x - x'|^2}{2\rho^2} + \delta_{ij}\sigma^2$$

Choice of covariance function

The **exponentiated quadratic (EQ) covariance kernel** assumes that the similarity between two observations will fall exponentially, i.e. that the rate of decline of the covariance increases with distance.

$$k(x, x') = \eta^2 exp - \frac{|x - x'|^2}{2\rho^2}$$

here, η^2 reflects the maximum covariance between any observed data points x, x', also known as the signal variance; $|x - x'|$ is the distance between any points x, x' along our continuous dimension; and ρ is the length scale, the rate of change with distance.

The noise term σ needs to be added to the diagonal of the covariance matrix to ensure it remains positive definite. It can be either fixed at a small positive constant, or modelled as a hyperparameter itself, to account for "noise variance" or "observation variance" (Rasmussen & Williams, 2005).

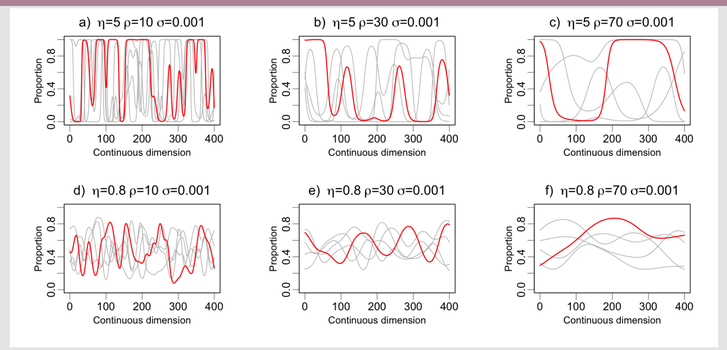

Figure 2. Simulated realisations from Gaussian Process Beta Regression where the distance |x − x′| is defined across a single dimension. Each grey line is an individual simulation (5 total), with a single simulation highlighted in red. Note how adopting the beta distribution successfully restricts the predictions to the standard unit interval (0,1).

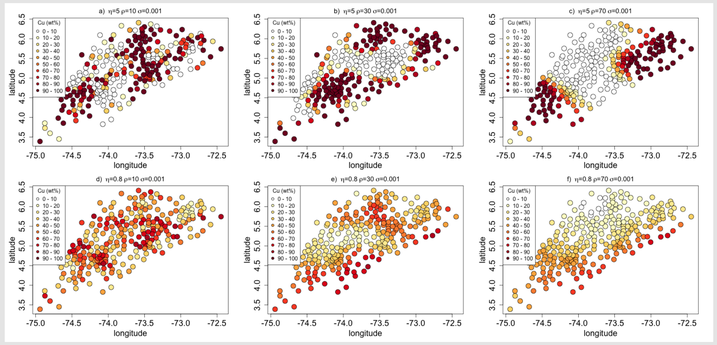

Figure 3. Simulated realisations from Gaussian Process Beta Regression, where the distance |x − x′| is defined across space, i.e. the two dimensions of latitude and longitude.

Gaussian Process Priors

In a Gaussian Process (GP), the Gaussian distribution is generalised to infinite dimensions to estimate the covariance between input-output sets (in our case, **input** being the distance between two artefact recovery locations, and **output** being the random intercept parameter describing the deviations of copper proportions from the global baseline values in the recovered artefacts). In formal terms, a GP is defined as:

$$f(x) \sim GP(m(x), k(x, x'))$$

where the function of x is distributed according to a Gaussian process with a mean function $m(x)$ and a covariance function $k(x, x')$ (Rasmussen & Williams, 2005). The mean function is often modelled as zero, as is the case here. This allows for focusing on the covariance function only, i.e. estimating the deviation of the Cu contents from the globally expected average Cu contents through space.

There are an infinite number of functions $f(x)$ that can fit some observed data points, and the goal is to find the functions that are the most consistent with the observed data through Bayesian posterior estimation.

These simulated examples show how Gaussian Process beta regression models can successfully capture both the rate and magnitude of change in proportions across space.

Limitations and future directions:

- To capture spatial variation in the rates and magnitudes of change, the GPR covariance functions could be extended to be non-stationary (e.g. Finley et al., 2019).
- In assuming the input data to be based on a continuous dimension, Gaussian processes are better suited to modelling point than polygon data. **However**, a GP can readily interpolate through space, whereas solutions such as ICAR estimate spatial correlation in terms of neighbourhoods — hence requiring aggregation to larger spatial units where the data of interest is missing for some polygons at the desired spatial resolution.

Methods and References

All simulations were coded using *Stan* (Stan Development Team, 2021), using the *RStan* package version 2.26.22 (Stan Development Team, 2021) as an interface to communicate between Stan and R, with all post-sampling analyses and graphs conducted in *R* v. 4.3.1 (R Core Team, 2023), and using *RStudio* v. 2023.02.2 (RStudio Team, 2023).

Ferrari, S., & Cribari-Neto, F. (2004).Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., & Banerjee, S. (2019).Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes. *Journal of Computational and Graphical Statistics*, *28*(2), 401–414. https://doi.org/10.1080/10618600.2018.1537924

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press. Retrieved January 31, 2023, from https://doi.org/10.7551/mitpress/3206.001.0001

Tobler, W. R. (1970).A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.