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

Funded by the European Union

European Research Council
Established by the European Commission

REVERSEACTION
REVERSE ENGINEERING COLLECTIVE ACTIONS

Spatial proportion data

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**: The bounded nature of proportion data often results 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 such an approach *can* alleviate these problems, it fails to account for the inconsistent variances of proportional response variables properly. 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.

Gaussian Process Beta Regression

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

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

$$\begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_{\text{location}} \end{pmatrix} \sim \text{GP}(0, K)$$

$$K_{ij} = \eta^2 \exp\left(-\frac{|x_i - x_j|^2}{2\rho^2}\right) + \delta_{ij}\sigma^2$$

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.

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

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 (fixed at a small positive constant, as observation variance in the model is captured by ϕ) is added to the diagonal of the covariance matrix to ensure it remains positive definite.

Methods and references

Please visit the GitHub webpage behind the QR code for details of the methods and bibliographic references.

Acknowledgments

This poster is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101021480, REVERSEACTION project). Funding was also obtained from a Cambridge AHRC-Doctoral Training Partnership Scholarship (2112128) and the Osk. Huttunen Foundation. We thank the Museo del Oro, Banco de la República, for their ongoing research collaboration, and the Museo del Oro and Carl M. Rodríguez for the photographs of the archaeological artefacts.

Case study

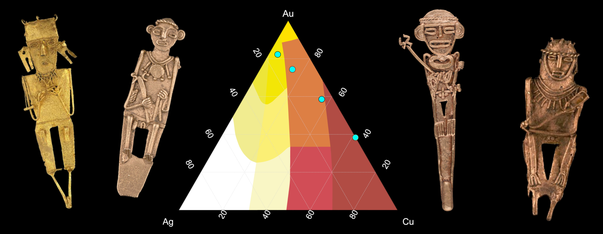

Figure 1. Examples of archaeological metal artefacts from the Muisca region of pre-Hispanic Colombia (AD 600-1600). The ternary diagram maps their chemical compositions onto the Au-Ag-Cu system, in other words, showing the relative proportions of their three main alloying constituents. Note the colour changes in the alloys based on their copper contents, shown in increasing order from left to right, with colour symbolism playing an important role for the region's inhabitants.

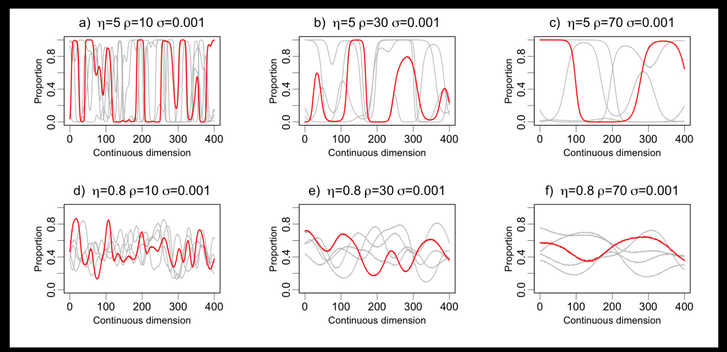

Figure 2. Simulated realisations from Gaussian Process Beta Regression where the distance |x − x′| is defined across a single dimension. Each grey line is an individual simulation (5 total), with a single simulation highlighted in red. Note how adopting the beta distribution successfully restricts the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

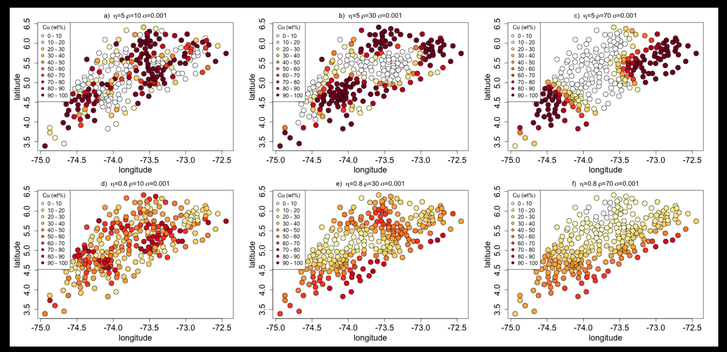

Figure 3. Simulated realisations from Gaussian Process Beta Regression, where the distance |x − x′| is defined across space, i.e. the two dimensions of latitude and longitude, with hypothetical examples of how the rate and magnitude of change in alloying recipes vary across space. Coordinate data from Departamento Administrativo Nacional de Estadística (2020) and Datos Abiertos Bogotá (2020).

Example data: chemical compositions of archaeological artefacts

Archaeologists often use chemical compositions of metal artefacts to understand which factors influenced alloy selection by craftspeople in the past. For instance, people could have intentionally manipulated the colour, melting point, or malleability of their alloys by changing their copper contents (see ***Figure 1***) — as informed by their socio-cultural contexts of metal production and consumption, and by the environmental availability of raw materials. To better understand how different groups of people varied in their technological practices, we can use archaeological sites as a proxy for these groups. **To model spatial signatures, we use inter-site distances as the input, and a varying intercept parameter as the output (describing changes in baseline compositions) in the Gaussian Process (GP) function.**

Social interpretation

For each hypothetical scenario, it is useful to compare the wiggliness of the simulations in 1D space in ***Figure 2*** to the changes in alloy compositions in ***Figure 3***.

- As we move from left to right, the distances at which alloy compositions change are increasingly larger, reflecting increasing values of ρ (the length scale).
- As we move between rows, the maximum covariance in alloy compositions changes, reflecting decreasing values of η and suggesting more similarity of compositions overall.
- For instance, ***Figure 3c*** has only three broad compositional groups, two dark red (copper-rich) and one white (low levels of copper), implying drastic changes in technological practices across sharp geographical boundaries (in real life, potentially corresponding to geographical barriers, such as mountain ranges, which can limit cross-cultural interaction and/or material exchange). In contrast, ***Figure 3f*** suggests a gradual change over a more restricted range of possibilities. This could imply the sharing of knowledge and resources between close neighbours, with some long-distance spatial process (e.g. availability of raw materials through exchange) influencing broader scale patterning in the background.

These simulated examples show how Gaussian Process beta regression models can successfully capture both the rate and magnitude of change in proportions across space. Applying these models to observed data allows obtaining probabilistic estimates of the GP functions that are the most consistent with the data. By accounting for sample interdependence arising from shared influences across space, we simultaneously obtain statistically more robust inferences.

Limitations and future directions

- We have previously proposed the adoption of variable dispersion sub-models within the beta regression framework, to examine the degree of standardisation in technological practices (Vieri, 2023; Vieri et al., 2024). To model how such dispersional signatures vary across space, we propose the adoption of GP priors for the linear ϕ submodel.
- GP priors can also be used to model selected covariates and how their impact on proportions varies through space.
- To capture changes in more than two proportional constituents simultaneously, a Dirichlet-distributed (Maier, 2014) GP model is proposed.
- Non-stationary GP covariance functions could be adopted to capture spatial variation in the rates and magnitudes of change (e.g. Finley et al., 2019).