Geographical Gaussian Process GAMs: a better spatially varying coefficient model than multiscale GWR?

Alexis Comber*[1], Paul Harris†[2] and Chris Brunsdon‡[3]

[1]School of Geography, University of Leeds, UK
[2]Sustainable Agriculture Sciences, Rothamsted Research, North Wyke, UK
[3]National Centre for Geocomputation, Maynooth University, Ireland

GISRUK 2023

Summary

This paper proposes a novel spatially varying coefficient model for spatial regression using General Additive Models (GAMs) with Gaussian Process (GP) splines parameterised with observation locations. The brand leader in this area is probably Multiscale GWR (MGWR) models but these have a number of theoretical and technical limitations. Here, a GAM with GP spline model and a MGWR model were applied to simulated spatial datasets with varying degrees of spatial autocorrelation. The GAM was shown to perform better than MGWR under a range of fit metrics. Some unresolved issues are discussed such as model calibration or tuning of knots and spline smoothing parameters.

KEYWORDS: Regression; GWR; GAM; Spatial Statistics.

1 Introduction

Linear regression seeks to model the relationship between a response variable, y, and a series of predictor variables (x_1, x_2 etc). It estimates a single set of unknown regression coefficients for each predictor variable (as well as independent error terms for each observation) by trying to minimise some metric such as the sum of the squared difference between the observed and predicted dependent variable (ordinary least squares). The standard form for a linear regression model is:

$$y_i = \beta_0 + \sum_{m}^{k=1} \beta_k x_{ik} + \epsilon_i \tag{1}$$

where for observations indexed by $i = 1 \ldots n$, y_i is the response variable, x_{ik} is the value of the

*a.comber@leeds.ac.uk
†paul.harris@rothamsted.ac.uk
‡christopher.brunsdon@mu.ie

k^{th} predictor variable, m is the number of predictor variables, β_0 is the intercept term, β_k is the regression coefficient for the k^{th} predictor variable and ϵ_i is the random error term.

However, in reality, the relationship between the response and predictor variables may change with observation location, potentially due to unaccounted for local factors or due to process spatial heterogeneity. Because of this 'whole map' or global regression models may unreasonably assume spatial stationarity in regression coefficients (Openshaw, 1996) and spatially varying coefficient regression models (SVCs) may be considered. An SVC model is one in which the regression coefficient estimates are allowed to vary over space, describing different predictor-to-response relationships in different locations:

$$y_i = \beta_0(u_i, v_i) + \sum_{m}^{k=1} \beta_k(u_i, v_i)x_{ik} + \epsilon_i \tag{2}$$

where now (u_i, v_i) are the spatial coordinates of the observations i and $\beta_k(u_i, v_i)$ are the coefficients estimated at those locations.

Spatially varying coefficient models are increasingly popular because their outputs provide intuitive measures of the scale of individual predictor-to-response relationships (i.e. how processes vary spatially) through determination of parameter specific bandwidths and the spatially varying coefficient estimates can be mapped (i.e. where processes vary), with values on similar scales to those estimated a standard OLS regression and thus easily understood by users. A popular approach is to solve Equation 2 with a geographically weighted regression (GWR) model (Brunsdon et al., 1996) or a multiscale GWR (MGWR) model (Yang, 2014; Fotheringham et al., 2017; Oshan et al., 2019), which is now recommended as the default GWR (Comber et al., 2022).

However there are number of conceptual limitations to geographically weighted approaches: observations are used in multiple local regression models rather than a single one; they violate standard assumptions of standard independent and identically distributed statistical error terms; they are also sensitive to the kernel form (shape) used to weight observations and different conclusions may be drawn about the effect of covariates under different kernel forms; they may be sensitive to local collinearity even when this is not a problem across the whole dataset (Wheeler and Páez, 2010; Páez et al., 2011); they are difficult to use for predictions at locations with no observation (Fan and Huang, 2022). As a result some argue that GWR and MGWR models are best suited for exploratory spatial analyses (Wheeler and Calder, 2007; Farber and Páez, 2007) and to guide further investigation through enhanced through process understanding (Comber et al., 2022).

Generalised Additive Models (GAMs) with Gaussian Process splines parameterised with location offer an alternative and novel approach to solve Equation 2 in order to to support process spatial understanding, to quantify process spatial heterogeneity and to support spatial prediction. This paper describes a multiscale spatially varying coefficient modelling using a Geographical Gaussian Process GAM (GGP-GAM).

2 Geographical Gaussian Process GAM

An alternative approach to solving Equation 2 is to use Gaussian Processes (GPs) to model terms in a Generalised Additive Model (GAM) (Wood, 2006; Fahrmeir et al., 2021). A GP is a random process over functions and GAMs calibrate regression models with unspecified functions of the predictor variables, of the form:

$$y = \alpha + f_1(z_1) + f_2(z_2) + \cdots + f_m(z_m) + \epsilon \tag{3}$$

where z_j may be a vector.

These can be extended so that each $f_j(z_j)$ is a linear regression coefficient on another predictor x_j:

$$y = \alpha(z_0) + x_1 f_1(z_1) + x_2 f_2(z_2) + \cdots + x_m f_m(z_m) + \epsilon \tag{4}$$

If $z_0 = z_1 = \cdots z_m = z$ say, and z is a vector specifying spatial locations then this specifies a spatially varying coefficient model:

$$y = \alpha(z) + x_1 f_1(z) + x_2 f_2(z) + \cdots + x_m f_m(z) + \epsilon \tag{5}$$

One way of specifying $\alpha(z) \cdots f_m(z)$ is that each function is generated from a GP and each function estimate is an *a posteriori* estimate of a GPs with a zero mean. GPs also have a covariance function:

$$\kappa_m(\delta) = Cov(f_m(\delta), f_m(z + \delta)) \tag{6}$$

This controls the 'smoothness' of $f_m(z)$: as $\kappa_m(\delta)$ reduces, δ increases and the 'smoother' $f_m(z)$ tends to be. The GAM estimates parameters in each $\kappa_j(\delta)$ in order to estimate $f_m(z)$. In this way a GAM uses smooth functions of the predictor variables in which the values of y are assumed to be of an exponential distribution, such as a Gaussian one. If

$$y = f(x) + \epsilon \tag{7}$$

where f is the function being sought in the model, then in GAMs, rather than assuming y to be some linear function of x, a space of functions, or basis, is chosen of which f is some element. This allows the basic formula above to be expanded:

$$y = f(x) + \epsilon = \sum_{j=1}^{d} \beta_j(x)\gamma_j + \epsilon \tag{8}$$

where each β_j is a basis function of the transformed x and the γ are the corresponding regression coefficient estimates. One example of a basis is a Gaussian Process basis. If there are n distinct geographical locations in the data set, then knowing the locations and the covariance function κ allows the variance covariance function of the values of β_j in each location to be found, giving the variance covariance matrix R. This can be translated into a set of n basis vectors $\beta_j(x)$ (Hefley et al., 2017), and the GAM can be calibrated in this way. Thus the predictors in a GAM include smooth functions of some or all of the covariates, which allow for non-linear relationships between the predictors and the target variable.

3 A Simulation case study

Simulated spatial data sets with varying degrees of spatial heterogeneity were used to examine the performance of GGP-GAM and to compare that with the performance of a standard MGWR. The simulated data were created following Fotheringham et al. (2017) and used subsequently by others (e.g Fan and Huang (2022)), with the aim of simulating the coefficient estimates (β's) for Equation 9:

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + \beta_2(u_i, v_i)x_{i2} + \epsilon_i \tag{9}$$

Three surfaces were created with varying degrees of spatial heterogeneity over consider a 25 × 25 regular square grid. Each of the surfaces was assigned to the coefficients of the three predictor variables as follows:

$$\beta_{zero} = \beta_0 = 3 \tag{10}$$

$$\beta_{low} = \beta_1 = 1 + \frac{1}{12}(u + v) \tag{11}$$

$$\beta_{high} = \beta_2 = 1 + \frac{1}{324}\left[36 - (6 - \frac{u}{2})^2\right]\left[36 - (6 - \frac{v}{2})^2\right] \tag{12}$$

As with Fotheringham et al. (2017), the values for x_1 and x_2 were generated from a normal distribution in the range [0, 1], ϵ from normal distribution in the range [0, 0.25], and 50 surfaces were generated. The simulated true regression coefficient surfaces are shown in Figure 1.

Gaussian Process splines parameterised with observation location (i.e. a GGP-GAM) can be used within a GAM model. The `mgcv` R package (Wood and Wood, 2015) was used to construct the GAM with GP splines with a GP smooth and for the simulation data, the X and Y locations were extracted from (u, v) respectively. The splines optimise a parameter which controls the degree of smoothing of the data and as such potentially indicates the locally varying nature of the coefficient estimate in a similar way to MGWR bandwidths. The GPs modelled in the GAM function all have

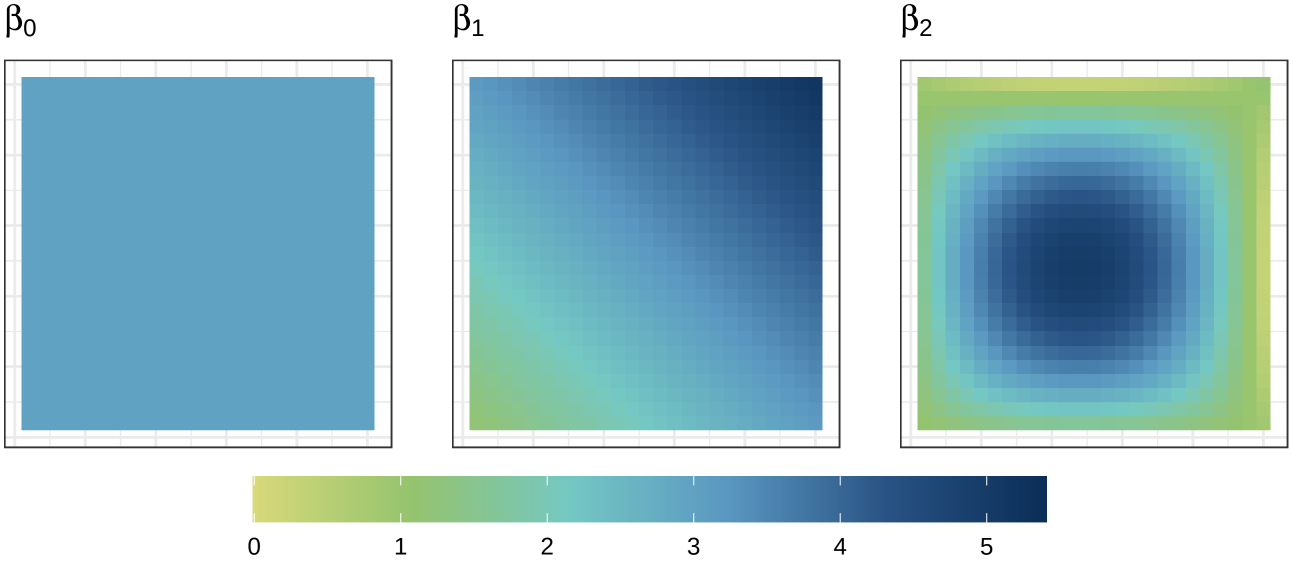

Figure 1: The simulated regression coefficient surfaces with zero, low and high spatial heterogeneity.

a mean of zero, so for each covariate an extra fixed offset term is added along with the spatially smoothed terms.

The simulated coefficient surfaces for β_0, β_1, β_1 in Figure 1 and the 50 simulated values of x_1, x_2 and ϵ were used to create a 50 sets of target variables over the 625 surface points, y. These values of true y, x_0, x_1 and x_2 were then used as inputs to the GGP-GAM model (i.e as 625 located observations with 4 fields) to generate coefficient estimates. For comparison a Multiscale GWR was also undertaken with the same data. The spatially located coefficient estimates for both models were retained and used to generate measures of fit by comparing the modelled coefficients with true ones in Figure 1, generating R^2 for β_1 and β_1 (as β_0 is stationary R^2 cannot be computed) and RMSE and MAE for β_0, β_1, β_1. This was done for each of the 50 sets of y, x_0, x_1 and x_2.

The results are shown in Figure 2. Under each fit measure, (AIC, RMSE, MAE, R^2) the GGP-GAM generates better estimates of the true coefficients than MGWR, with the difference in fit measures increasing with increasing degrees of spatial heterogeneity. This is also shown visually using an example set of β values from the 50 sets of coefficients estimated by the GGP-GAMand by the MGWR to recreate the surfaces for β_0, β_1, β_1. Figure 3 shows the coefficient surfaces estimated from the 10th set of simulations, with the same shading breaks as Figure 2. The better performance of the GGP-GAM in estimating the true βs is clearly demonstrated.

4 Conclusions

This paper demonstrates for the first time the application of a GGP-GAM model, through GAMs (Wood, 2006; Fahrmeir et al., 2021) with GP splines parameterised with observation location. Using simulated data with known spatial heterogeneity, the GGP-GAM models out-performed MGWR.

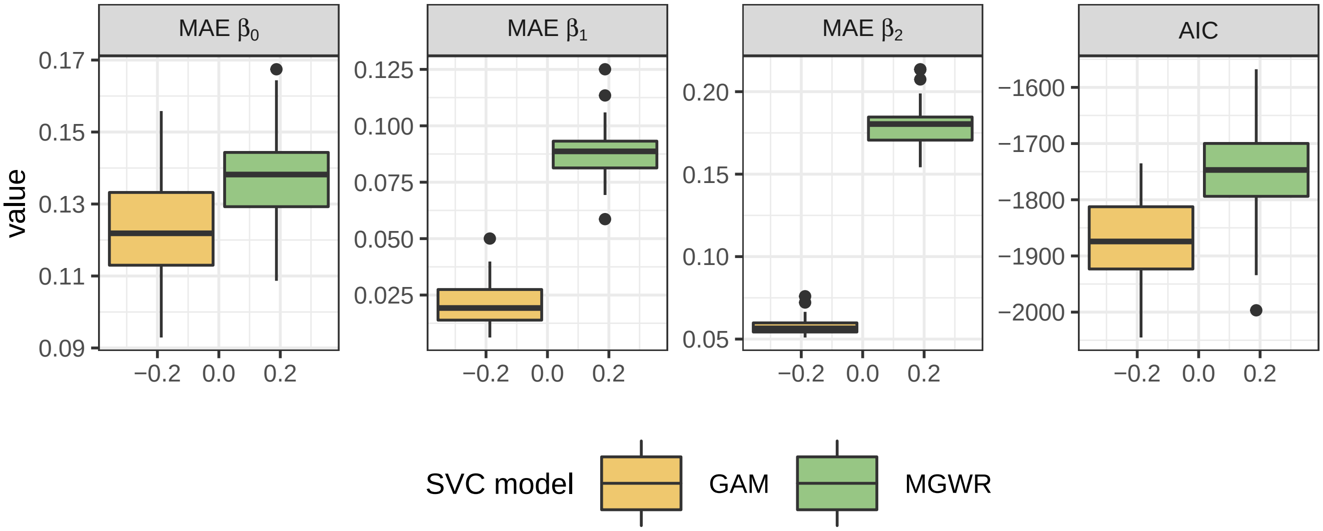

Figure 2: Evaluation of the accuracy of the GGP-GAM and MGWR regression coefficient estimates, when compared to the true coefficients.

GAMs offer an intuitive approach to fit relatively complex relationships in data with complex interactions and non-linearities. The outputs provide easily understood measures of the process spatial heterogeneity, the models predict well and support inferences about such relationships.

5 Biography

Lex Comber is Professor of Spatial Data Analytics, with research interests in all areas of spatial analysis and geocomputation. This year is he is mostly interested in methods for handling spatial scale and for supporting different scales of decision making.

Paul Harris is Professor of Agriculture Sciences with interests in environmental monitoring and modelling, and methodologies in spatial statistics. This year he is *still* mostly developing analytical methods for resilient farming systems.

Chris Brunsdon is Professor of Geocomputation at the National University of Ireland, Maynooth. He has research interests in developing spatial data analysis methods, and reproducibility and inference for these methods. Recently he has mostly been applying these ideas to data related to Covid-19.

References

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, *28*(4), 281–298.

Comber, A., et al. (2022). A route map for successful applications of geographically weighted regression. *Geographical Analysis*.

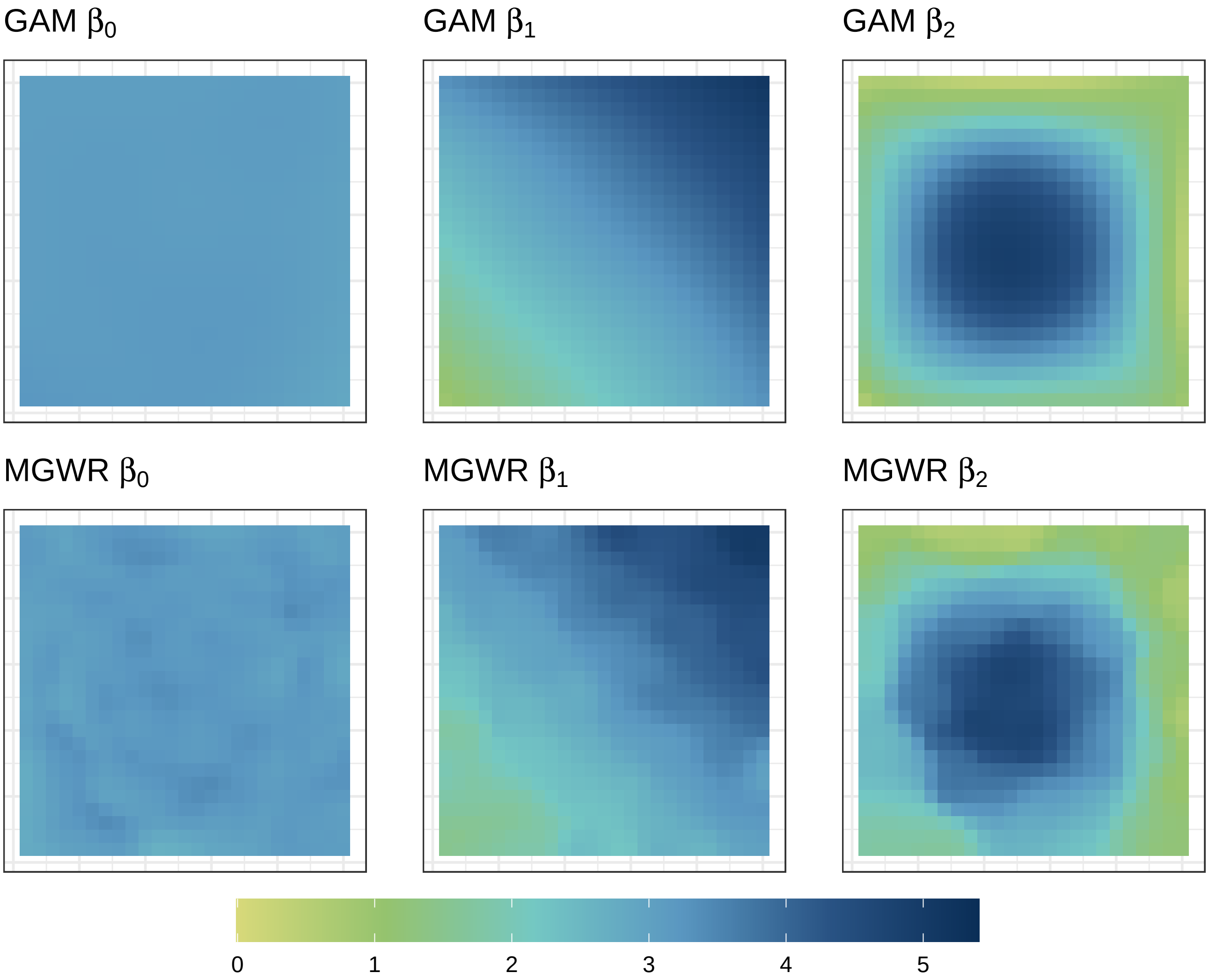

Figure 3: The estimated regression coefficients from a single sample of the GGP-GAM and MGWR models, shaded using the same range as Figure 1.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). Regression models. In *Regression* (pp. 23–84). Springer.

Fan, Y.-T. & Huang, H.-C. (2022). Spatially varying coefficient models using reduced-rank thin-plate splines. *Spatial Statistics*, *51*, 100654.

Farber, S. & Páez, A. (2007). A systematic investigation of cross-validation in gwr model estimation: empirical analysis and monte carlo simulations. *Journal of Geographical Systems*, *9*(4), 371–396.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, *107*(6), 1247–1265.

Hefley, T. J., et al. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, *98*(3), 632–646.

Openshaw, S. (1996). Developing gis-relevant zone-based spatial analysis methods. *Spatial analysis: modelling in a GIS environment*, (pp. 55–73).

Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Fotheringham, A. S. (2019). mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, *8*(6), 269.

Páez, A., Farber, S., & Wheeler, D. (2011). A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A*, *43*(12), 2992–3010.

Wheeler, D. C. & Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, *9*(2), 145–166.

Wheeler, D. C. & Páez, A. (2010). Geographically weighted regression. In *Handbook of applied spatial analysis* (pp. 461–486). Springer.

Wood, S. & Wood, M. S. (2015). Package 'mgcv'. *R package version*, *1*(29), 729.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman Hall/CRC.

Yang, W. (2014). *An extension of geographically weighted regression with flexible bandwidths*. PhD thesis, University of St Andrews.