The Journal of Open Source Software

stgam: An R package for GAM-based varying coefficient models

Lex Comber [1*], **Paul Harris** [2*], and **Chris Brunsdon** [3*]

1 School of Geography, University of Leeds, UK **2** Rothamsted Research, North Wyke, UK **3** National Centre for Geocomputation, Maynooth University, Ireland * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

We are often interested in understanding how and where statistical relationships vary over space, and how they change over time. Quantifying such *process heterogeneity* (spatial and or temporal) can be done using *varying coefficient* models. The opportunity to undertake such space-time analyses are greater due to the increased generation and availability of data that include both spatial and temporal attributes (e.g. GPS coordinates and time-stamps). The **stgam** package provides a framework for creating regression models using Generalized Additive Models (GAMs) (Hastie & Tibshirani, 1990) in which the relationships between the response (dependent) variable and individual predictor (independent) variables are allowed to vary over space, time or both. It addresses a key question of the form space-time interaction to be specified in models. Frequently randomly guessed have to be made about the form of predictor-to-response relationships until an effective model form and associated model fits are found. To address this the **stgam** package includes functions to create multiple models, each specifying different relationships between the response variable y and the predictor variables $x_1 \dots x_n$. Each model is evaluated using the Bayesian Information Criterion (BIC) (Schwarz, 1978) which approximates the likelihood (probability) of the model being the correct model, given the data used in the analysis. Where multiple models are highly probable, then these can be be combined using a Bayesian Model Averaging approach (Brunsdon et al., 2023; Fragoso et al., 2018). Finally the **stgam** package contains functions for creating spatially and / or temporally varying regression coefficient estimates. These can be mapped in the usual way to show how where and when the relationships between the response and the predictor variables vary over space and time.

Statement of need

A number approaches have been established to quantify process spatial non-stationarity or heterogeneity (Brunsdon et al., 1996; Casetti, 1972; Fotheringham et al., 2002; Gelfand et al., 2003; Griffith, 2008; Jones, 1991; McMillen, 1996) and tools also exist to quantify the *temporal* dynamics of these (Crespo et al., 2007; Di Giacinto, 2006; Elhorst, 2003; Gelfand et al., 2004; Huang et al., 2010; Pace et al., 1998, 2000). All existing models require the user to make decisions about the nature of the space-time relationships in the data and thus the model and assume the presence of latent spatial and temporal autocorrelation in variables. The most commonly used approach in geographical analyses of space-time problems is geographically and temporally weighted regression (GTWR) (Huang et al., 2010) which seeks to optimises a single space-time kernel to define the space-time relationships of all covariates with the target variable. Some recent steps in the right direction have been taken in this regard: Liu et al. (2017) developed a semi-parametric temporal extension to mixed geographically weighted regression, in which some relationships and coefficients are assumed to be globally constant and others vary locally over time; Hong et al. (2021) used a bootstrap approach to identify global coefficients in such models, but still define the same space-time relationship for each varying covariate.

To address this gap, the **stgam** package draws from the `mgcv` GAM package (S. N. Wood, 2017). GAMs are able to handle many kinds of responses (Fahrmeir et al., 2021) and provide an intuitive approach to fit relatively complex relationships in data with complex interactions and non-linearities (S. N. Wood, 2017). The outputs of GAMs provide readily understood summaries of the relationship between predictor and response variables, and how the outcome is modelled. They are able to predict well from complex systems, quantify relationships, and support inferences about these relationships, such as which variables are important and at what range of values they are most influential (Hastie & Tibshirani, 1990; S. N. Wood, 2017). GAMs can perform as well as or better than most machine learning models and they are relatively fast (Friedman, 2001; S. N. Wood, 2017). Importantly, in the context of varying coefficient modelling, GAMs combine predictive power, model accuracy and model transparency. GAMs model non-linear relationships using smooths, also referred to as splines. These can be of different forms depending on the problem and data (Hastie & Tibshirani, 1990) and, as they can be represented as linear combinations of basis functions, they are sometimes referred to as Gaussian Process (GP) splines. A GP is a random process over functions, and its terms can be modelled using GP-splines within a GAM. Thus GP splines are represented as a linear combination of non-linear *basis functions* of predictor variables, which can generate predictions of the outcome variable. Basis functions can be either single or multi-dimensional in terms of predictor variables. As a result, a GAM consists of linear sums of multi-dimensional basis functions that allow complex relationships to be modelled. The appropriate degree of "wiggliness" in each spline is determined by the smoothing parameters, which balances over-fitting versus capturing the complexity in the relationship. GAMs with GP smooths parameterised with location and / or time can be used to construct regression models that allow coefficient estimates to vary and capture process spatial and / or temporal heterogeneity: a varying coefficient approach. While spline-based varying coefficient models have been proposed, for example using a generalized linear model with reduced-rank thin-plate splines (Fan & Huang, 2022), here the predictor-to-response relationship is considered over space and time as a GP.

A final consideration is the need to determine model form. Standard approaches for constructing spatially and temporally varying coefficient models, which in the absence of a theoretical model, commonly assume that some degree of spatial and temporal dependence *is* present in the data and in the relationship of each covariate with the target variable. In the GAM GP smooth approach, each covariate would be specified in a SP smooth parameterised with location and with time under the assumption that any temporal trends in coefficient estimates *will* vary with location. However, each covariate can be specified in six different ways: it can be omitted, included as a parametric (global) term, in a smooth with location, in a smooth with time, in a smooth with location *and* time, or in two separate space and time smooths. The last five options similarly apply to the intercept.

Package overview

The **stgam** package contains functions to support varying coefficient modelling using GAMs with GP smooths, that provide a wrapper for the GAM implementation in the `mgcv` package (S. Wood, 2015), that create, evaluate, and aggregate multiple models. It also contains two datasets that are used to illustrate the functions. These are described in Table 1.

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

Brunsdon, C., Harris, P., & Comber, A. (2023). *Smarter than your average model? Bayesian model averaging as a spatial analysis tool*. 12th International Conference on Geographic

Table 1: Spatial models currently implemented in **geostan**.

	Gaussian	Student's t	Poisson	Binomial
CAR	x		x	x
ESF	x	x	x	x
GLM	x	x	x	x
ICAR			x	x
SAR	x		x	x

Information Science.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographical research. *Geographical Analysis*, *4*(1), 81–91.

Crespo, R., Fotheringham, S., & Charlton, M. (2007). Application of geographically weighted regression to a 19-year set of house price data in london to calibrate local hedonic price models. *Proceedings of the 9th International Conference on Geocomputation*.

Di Giacinto, V. (2006). A generalized space-time ARMA model with an application to regional unemployment analysis in italy. *International Regional Science Review*, *29*(2), 159–198.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review*, *26*(3), 244–268.

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.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. John Wiley & Sons.

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, *86*(1), 1–28.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Gelfand, A. E., Ecker, M. D., Knight, J. R., & Sirmans, C. (2004). The dynamics of location in home price. *The Journal of Real Estate Finance and Economics*, *29*, 149–166.

Gelfand, A. E., Kim, H.-J., Sirmans, C., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, *98*(462), 387–396.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, *40*(11), 2751–2769.

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. Chapman hall & CRC. *Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 1*.

Hong, Z., Mei, C., Wang, H., & Du, W. (2021). Spatiotemporal effects of climate factors on childhood hand, foot, and mouth disease: A case study using mixed geographically and temporally weighted regression models. *International Journal of Geographical Information Science*, *35*(8), 1611–1633.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, *24*(3), 383–401.

Jones, K. (1991). Specifying and estimating multi-level models for geographical research.

Transactions of the Institute of British Geographers, 148–159.

Liu, J., Zhao, Y., Yang, Y., Xu, S., Zhang, F., Zhang, X., Shi, L., & Qiu, A. (2017). A mixed geographically and temporally weighted regression: Exploring spatial-temporal variations from global and local perspectives. *Entropy*, *19*(2), 53.

McMillen, D. P. (1996). One hundred fifty years of land values in chicago: A nonparametric approach. *Journal of Urban Economics*, *40*(1), 100–124.

Pace, R. K., Barry, R., Clapp, J. M., & Rodriquez, M. (1998). Spatiotemporal autoregressive models of neighborhood effects. *The Journal of Real Estate Finance and Economics*, *17*, 15–33.

Pace, R. K., Barry, R., Gilley, O. W., & Sirmans, C. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, *16*(2), 229–246.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Wood, S. (2015). Package "mgcv." *R Package Version*, *1*(29), 729.

Wood, S. N. (2017). *Generalized additive models: An introduction with r*. CRC press.