Using GAMs to understand whether and how processes vary over space and time

Comber A*[1], Harris P†[2] and Brunsdon C‡[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 2025

Summary

This paper an informed approach for constructing varying coefficient regression models using Generalized Additive Models (GAMs) with Gaussian Process (GP) smooths. Using a house price data over 13 years for a local case study, it investigates different model forms in order to determine the most probable model given the data. It determines the presence of any space-time dependencies between the target and each predictor variable is present, and if so their nature (ie whether they are independent or interact). It does this to avoid assumptions about the nature of spatial and / or temporal dependencies, in contrast to many existing approaches for space-time regressions which implicitly include baked in assumptions assumptions about the presence of these. The analysis uses tools in the `stgam` R package to undertake the analysis and to extract the varying coefficients from the model and a number of areas of further work are identified.

KEYWORDS: Space-time relationships, Coefficient non-stationarity, Process heterogeneity, Regression, Inference.

1 Introduction

Geographical analyses are frequently concerned with quantifying how, where and more recently when processes and statistical relationships vary over and space and / or over time. A common goal in regression models is to capture process heterogeneity (non-stationarity) through *local*, spatially varying coefficient (SVC) models, using approaches such as Geographically Weighted Regression (GWR) (Brunsdon et al., 1996) and multiscale GWR (Yang, 2014; Fotheringham et al., 2017) and recent improvements based on Generalised Additive Models (GAMs) (Comber et al., 2024b,a). In

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

the spatial case, maps of the spatial distribution of local coefficient estimates from a SVC model provide insights into the spatial variations in the relationship between the response variable the factors assumed to driving it, and can be used to inform policy or to guide further investigation.

This paper describes an extension of SVC models into the temporal domain and proposes a workflow for explicitly capturing the nature of any space-time dependencies (interactions), rather than making assumptions about them, as is currently done all other approaches to space time regression modelling used in geographical analyses (e.g. GTWR Fotheringham et al. (2015); **?**). It does this by extending the application of GAMs (Hastie, 2017) with Gaussian Process (GP) smooths described in Comber et al. (2024a) for SVC modelling, into space time modelling to create spatially and temporally varying coefficient (STVC) models. The use of GP smooths parameterised with location within GAM based SVCs, captures a key concept in geographical analysis, that of distance decay linked to spatial autocorrelation. This is reflected in Tobler's First Law of Geography '*everything is related to everything else, but near things are more related than distant things*' (Tobler, 1970, p 236). Parallel arguments can be made with respect to the temporal dimension (Comber and Wulder, 2019).

However a critical consideration is how to treat space and time, with options to specify each predictor variable in a number of different ways in the model, each making assumptions about spatial and temporal dependencies. GTWR for example, optimises a space time kernel, on the implicit assumption that the spatial and temporal processes **do** interact, and the temporal trends in predictor-target variable relationships *will* vary over space (as opposed to being independent of location). Thus a critical consideration in varying coefficient modelling is to determine how to specify the covariate terms within the model. This is straightforward if there is a clear theory defining the predictor-to-response interactions. However, data analyses is often used to explore data space-time interactions. Thus this paper suggests a workflow to underpin process understanding is the primary objective (i.e. to understand the nature of any spatial and / or temporal dependencies) rather than prediction. It does this using the frameworks in the `stgam` R package (Comber et al., 2024c) that supports explicit investigation of the nature of any spatial and / or temporal dependencies in the data, that identifies the most probable model form or set of model forms, and that allows the user to make spatio-temporal predictions over unsampled locations and / or time periods. The ability to construct robust varying coefficient models in this way is key in a number of critical research areas (for example, analysis of resilience to climate change, where the aim is to determine the varying drivers of changes in resilience and to predict tipping points).

2 Data and Methods

This study uses time series data of annual mean house price over 13 years (2008 to 2020) in 482 lower super output areas (LSOAs) in around Leeds. The data also contains annual data i) on the proportion of the population receiving Job Seeker's Allowance (JSA - an unemployment benefit), ii) Churn, residential mobility measure describing the proportion of LSOA population change, and iii) the proportion of the population registering for a new National Insurance Number (NINO) as a measure of in-migration. These form the predictor variables of house price in the model. The study area is shown in Figure 1.

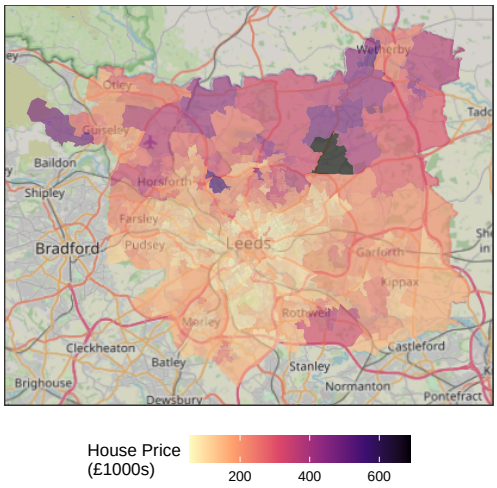

Figure 1: LSOAs in the study area and mean house price for 2017, with an OpenStreetMap backdrop.

The `stgam` R packages was used to undertake the varying coefficient modelling, which uses the `mgcv` package (Wood, 2015) to provide GAM with GP smooths functionality. It has at its core two key ideas: to make no assumptions about the presence of spatial and / or temporal dependencies in the data and the model specification, and to allow the user to evaluate and combine multiple highly probable models.

To illustrate the first of these, the code below creates a hypothetical GAM-based varying coefficient model that includes GP smooths parameterised with observation location (`X` and `Y`) and time (`Ti`) for each predictor variable. An intercept column is defined in the data to allow it to be treated as an addressable term, each term is assumed to take a spatio-temporal form in a smooth rather than a parametric form as in a standard OLS regression. Thus space-time dependencies are assumed to be present as in other STVC models such as GTWR:

```
# do not run!
gam.stvc.mod = gam(y ~ 0 +
                     s(X, Y, Ti, bs = "gp", by = Intercept) +
                     s(X, Y, Ti, bs = "gp", by = x1) +
                     s(X, Y, Ti, bs = "gp", by = x2),
                   data = data |> mutate(Intercept = 1))
```

However, each predictor variable could be specified in a number of different ways, for example with separate smooths for time and location, rather than the combined one as above. Therefore an important consideration is how best specify the model. For SVC models there are 3 possible ways that each predictor variable could be specified:

1. It is omitted.
2. It is included as a parametric response with no GP smooth.

3. It is included in a GP smooth parameterised with location.

For STVC models there are an additional 3 ways that each covariate can be specified:

4. It is included a GP smooth parameterised with time (but not location).
5. It is included in a single GP smooth parameterised with location and time.
6. It is included in 2 separate GP smooths, one parameterised with location and the other with time.

The intercept can be treated similarly, but without it being absent. Thus for a SVC regression model with k predictor variables there are 2×3^k potential models and for a STVC 5×6^k models.

In this case there are 1080 potential models to evaluate which can be done using a probability based approach. It can be shown that the model Bayesian Information Criterion (BIC) (Schwarz, 1978) of each model can be used to derive the likelihoods (probabilities) of each individual model M_i being the correct model, given the data D. Thus a BIC value for each potential model was calculated and used generate probabilities. This allowed different model forms to be ranked and compared.

3 Results

The 10 most probable models are shown in Table 1. The model probability ($Pr(M)$) indicates the uncertainty over model specification (model form). The relative probabilities in Table 1 indicate that there is 3% chance (0.031) that the second ranked model is better than the first, and a 2% chance (0.023) that the third ranked model is better than the first. In this case, the top ranked model can be selected. This indicates that the most probable model is one which specifies the predictor variables in the following way:

- the Intercept is included in separate space and time GP smooths, indicating space and time independence.
- Churn is not included (selected) in the model.
- NINO is is included only within a spatial smooth indicating spatial dependencies.
- JSA is included in separate space and time GP smooths, indicating space and time independence.

Perhaps not surprisingly for this local case study, none of the top 10 model forms in Table 1 indicate combined space-time GP smooths and thus space-time dependencies. This is because observations in a relatively small case study area are likely to be subject to the same socio-economic drivers, which in turn suggests that any space and time effects will be independent of each other. Whereas, in a national study for example, space and time effects might be expected to interact more strongly.

Summaries of the STVC model estimates and their variation over space and time are shown in Table 2. Interestingly, the NINO each predictor variable flips in sign at some point in space-time. It is instructive to examine some of this temporal and spatial variation as in Figures 2 and 3. The Intercept increases over time and NINO, with no temporal smooth, has consistent temporal

relationship with house price over time (these plots in Figure 2 show the annual mean). The relationship of JSA with house price is increasingly negative after 2014. The changes of the JSA coefficient estimates (Figure 3) indicate that these changes are relatively even over space.

Table 1: The 10 best models, and how the predictor variables were specified within the model, where '—' indicates the absence of a predictor, 'Fixed' that a parametric form was specified, 's_S' a spatial GP spline, 's_T' a temporal GP spline and 's_ST' a spatio-temporal GP spline.

Rank	Intercept	Churn	NINO	JSA	BIC	Pr(M)
1	s_T + S_S	—	s_S	s_T + s_S	64370.23	–
2	s_T + S_S	s_T	s_S	s_T + s_S	64377.11	0.031
3	s_T + S_S	Fixed	s_S	s_T + s_S	64377.74	0.023
4	s_T + S_S	s_T	s_T + s_S	s_T + s_S	64396.77	0.000
5	s_T + S_S	—	s_T + s_S	s_T + s_S	64405.02	0.000
6	s_T + S_S	Fixed	s_T + s_S	s_T + s_S	64412.94	0.000
7	s_T + S_S	s_S	s_S	s_T + s_S	64452.29	0.000
8	s_T + S_S	s_T + s_S	s_S	s_T + s_S	64460.64	0.000
9	s_S	s_T	s_S	s_T + s_S	64492.65	0.000
10	s_T + S_S	s_S	s_T + s_S	s_T + s_S	64495.16	0.000

Table 2: Summaries of the STVC coefficient estimates and their variation over space and time.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Intercept	98.0	152.3	183.7	200.6	233.2	540.5
NINO	-21172.6	-730.3	-130.3	-463.4	134.9	2390.6
JSA	-13634.2	-4191.0	-2380.5	-2974.3	-1088.0	-93.8

4 Acknowledgements

This research is supported by UKRI (ESRC) funding ES/Y006259/1 under the Digital Footprints scheme. The data for was assembled by Jennie Grey (`https://orcid.org/0000-0002-7558-1247`) during her ESRC funded PhD research (ES/P000401/1). Her thesis can be found at `https://etheses.whiterose.ac.uk/id/eprint/33039/`.

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 varying coefficient approaches methods and how they interact with spatial and temporal scale.

Professor Paul Harris is a Principal Research Scientist at Rothamsted Research. He has research interests in the development of agroecological modelling frameworks for sustainable and resilient agriculture.

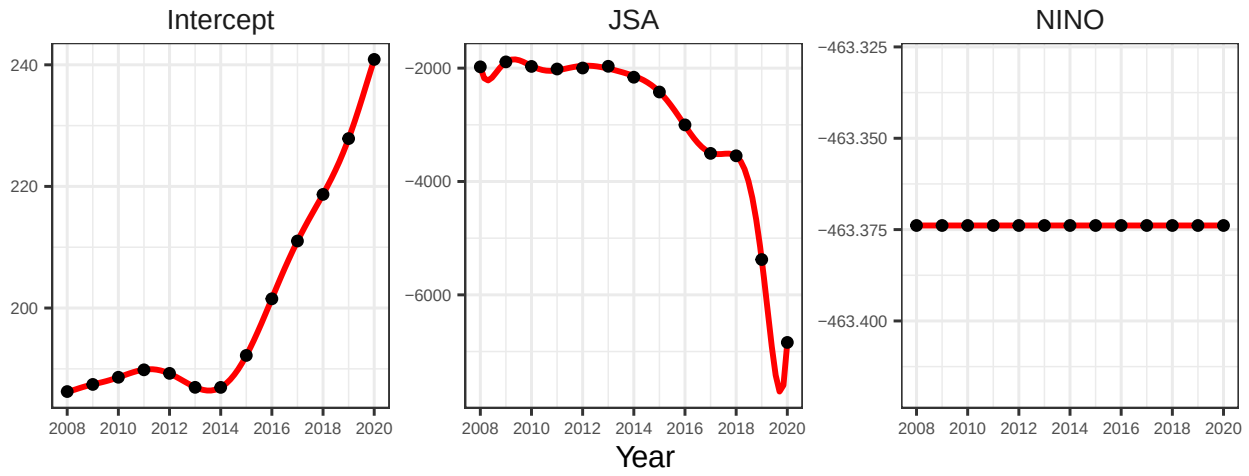

Figure 2: The temporal variation of the coefficient estimates, 2008 to 2020.

Chris Brunsdon is Professor of Geocomputation and Director of the National Centre for Geocomputation at the National University of Ireland, Maynooth. A researcher in spatial data science, he is best known for his work on geographically weighted regression (GWR) and other local statistical modeling techniques. He has co-authored influential texts in spatial analysis—and has contributed extensively to open-source software for geospatial research, particularly in the R environment.

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., Harris, P., & Brunsdon, C. (2024a). Multiscale spatially varying coefficient modelling using a geographical gaussian process gam. *International Journal of Geographical Information Science*, *38*(1), 27–47.

Comber, A., Harris, P., Murakami, D., Nakaya, T., Tsutsumida, N., Yoshida, T., & Brunsdon, C. (2024b). Encapsulating spatially varying relationships with a generalized additive model. *ISPRS International Journal of Geo-Information*, *13*(12), 459.

Comber, A. & Wulder, M. (2019). Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use.

Comber, L., Harris, P., & Brunsdon, C. (2024c). *stgam: Spatially and Temporally Varying Coefficient Models Using Generalized Additive Models*. R package version 0.0.1.0 `https://CRAN.R-project.org/package=stgam`.

Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Geographical and temporal weighted regression (gtwr). *Geographical Analysis*, *47*(4), 431–452.

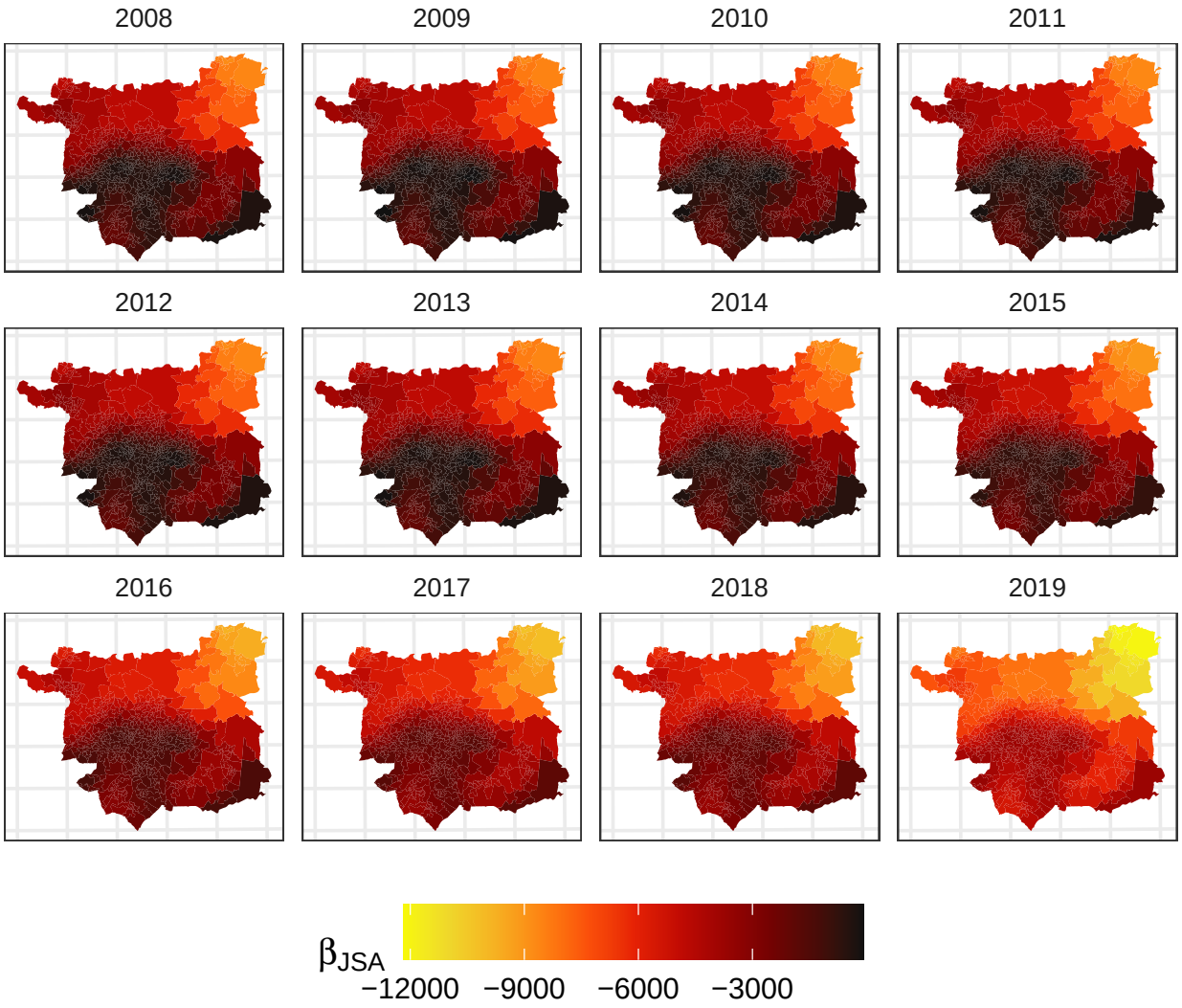

Figure 3: Detail of the spatial and temporal variation of the JSA coefficient estmates, 2008 to 2019, in Leeds.

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

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Routledge.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, (pp. 461–464).

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

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

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