Modelling Spatial Heterogeneity: a Local Approach or a Global Approach?

Guanpeng Dong[1], Richard Harris

[1]School of Geographical Sciences, University of Bristol, University Road, Bristol. BS8 1SS.

Tel. (+44) 117 928 9954

guanpengdong@gmail.com

KEYWORDS: spatial heterogeneity, GWR, spatially varying coefficients model

1. Introduction

Spatial heterogeneity, the local departures from the average trend, has recently attracted much attention when analysing geographical data in many research fields. Here we limit ourselves to two approaches designed to model spatial heterogeneity: geographically weighted regression (GWR) (Fotheringham et al. 2002) and Bayesian spatially varying coefficient model (SVC) (Gelfand et al. 2003).

GWR, in nature, is a local modelling approach in that it models spatial heterogeneity by sequentially fitting a series of weighted and localised regression models, centred upon each fit location. The weights are inversely proportional to the geographic distance between the fit points and the observations. SVC, on the other hand, is a global modelling approach in that it regards the spatial heterogeneity in regression coefficients as a multivariate Gaussian process and calibrates only one model rather than n localised models as in GWR.

Our intention is to illustrate the relative advantages of SVC over GWR in terms of conducting inferences on spatial heterogeneity and quantifying prediction uncertainty.

2. Methods and Data

2.1 GWR

A detailed description of GWR is given by Fotheringham et al. (2002), with a summary provided by Nakaya (2008). Following Wheeler and Calder (2007), the basic specification of GWR is,

$$y(s)=X(s)\beta(s)+\varepsilon(s) \qquad (1)$$

where s is a location indicator, y(s) and X(s) are the outcome and predictor variables at location s, β(s) are the estimated coefficients at location s, ε(s) is the random error and ε(s) ~ N(0, τ2). β(s) is given by $\beta(s) = [X^T W(s) X]^{-1} X^T W(s)y$. W(s) is a locally varying diagonal matrix defining the inverse distance weighting of the observations surrounding the fit location, s.

Several kernel functions could be used in specifying W(s) (Fotheringham et al. 2002). Here we employ the exponential kernel, W(s,s*) = exp (-φ×Ds,s*), where Ds,s* is the geographic distance between locations, s and s*, and φ is the spatial decay parameter, obtained using a cross-validation procedure.

2.2 SVC

Drawing upon Finley (2011), a standard SVC is specified as,

$$y(s)=X(s)\beta+\tilde{X}(s)w(s)+\varepsilon(s) \qquad (2)$$

where y(s), X(s) and ε(s) are the same as in GWR, and $\tilde{X}(s)$ refers to covariates from X(s) whose regression coefficients are assumed to be spatially varying. β now is the global coefficients for the covariates and spatially varying coefficients are obtained by β + w(s). Suppose we have p predictor variables including an intercept and n samples, then y is an n × 1 vector of outcome variable, X is an n × p design matrix, $\tilde{X}$ is an n × np block diagonal matrix specified through subsequently shifting each row in X by p places along with 0 in other places (Wheeler and Calder 2007).

The core of SVC is the specification of w(s), the spatial random effects for covariates. In SVC, w(s) = [w1(s), w2(s), ... , wp(s)] is assumed to follow a multivariate Gaussian process, w(s) ~ MVGP(0, Cw(s, s*)) where Cw(s, s*) = Cov(w(s), w(s*)) is a cross-covariance matrix function (Finley 2011). Different specifications of Cw(s, s*) lead to different covariance structures of w(s), Σw(s) and huge differences in computation burden (for the construction of more complex Cw(s, s*) based on the Linear Model of Coregionalisation see Finley, 2011).

In this study, we employ a more widely used separable SVC specification (Gelfand et al. 2003; Wheeler and Calder 2007; Wheeler and Waller 2009), where the np × np covariance matrix of w(s), Σw(s) = H(φ)⊗T. H(φ) is the n × n spatial correlation matrix among n locations determined by φ. Each entry of H(φ) is calculated by H(φ)[s, s*] = exp (-φ×Ds,s*). T is a positive definite p × p matrix capturing the covariance of the spatial random effects, with the diagonal elements being the spatial variances for them. ⊗ is the Kronecker product operator (Wheeler and Calder 2007). The SVC specification will be completed by further specifying the prior distributions for each unknown parameters (β, φ, T, τ2) in the model, and we implement this model according to the MCMC algorithm provided by Finley (2011, appendix A).

2.3 Synthetic data

The data generating process is Equation (2). We have three predictor variables in the synthetic data set, X = [1, X1, X2] where X1 and X2 are drawn from two independent standard normal distribution; ε(s) is drawn from a normal distribution with mean 0 and variance (τ2) 1.5. The global regression coefficients β are set to 5, 2 and 3. The multivariate Gaussian process of w(s) is generated by setting the spatial decay parameter φ = 0.6 and T = diag (2, 4, 8).

We randomly simulate 300 locations with both x and y coordinates within 0 to 10 units. Then a subset of 200 locations is randomly drawn to estimate models while the remaining 100 locations are served as a handout set to assess model prediction accuracy.

3. Results and Discussion

3.1 GWR and SVC estimate results

The SVC estimate results are based on one MCMC chain with random initial values, which run 25000 iterations. The CODA package in R is used to diagnose convergence and after

10000 iterations the chain is well mixing and convergence is diagnosed (Plummer et al. 2006). All the SVC results in Table 1 are based on the 15000 after burn-in posterior samples.

From Table 1, we can see that all the true values of the unknown parameters are within the 95% credible intervals of SVC estimates. The model fit is better in SVC than GWR (the root mean squared error (RMSE) for y is much lower in SVC than GWR), which is as expected because the true data generating process is specified according to SVC.

Table 1. Summarising estimation results from GWR and SVC.

	True values	GWR estimates	SVC estimates
Intercept	4.18 (1.54, 7.44)	5.92 (4.05, 8.52)	5.64 (3.64,7.58)
Beta1	2.93 (-1.00, 5.59)	3.28 (-0.09, 5.29)	3.22 (0.49, 5.85)
Beta2	3.93 (-0.00, 6.59)	3.86 (1.84, 5.73)	3.45 (0.93, 7.58)
φ	0.6	0.97	0.69 (0.37, 1.17)
T_{11}	2		1.42 (0.30, 2.96)
T_{22}	4		4.47 (2.53, 8.24)
T_{33}	8		4.83 (2.38, 9.53)
T_{21}	0		0.15 (-0.98, 1.39)
T_{31}	0		-0.27 (-1.68, 1.01)
T_{32}	0		0.99 (-0.89, 3.30)
τ^2	1.5	2.31	0.70 (0.17, 2.44)
RMSE		1.88	0.37

Note: Intercept, Beta1 and Beta2 refer to the coefficient surfaces, which equal to β + w(s)

The estimated surfaces for three regression coefficients from GWR and SVC, and the true surfaces are illustrated in Figure 1. It clearly shows that the surfaces estimated by GWR are much smoother than that estimated by SVC and the true surfaces, indicating the potential problem of over-smoothing for GWR, as noted by Wheeler and Waller (2009).

3.2 Conducting inference on spatial heterogeneity in SVC

Often we are interested in diagnosing where the local coefficients are significantly different from zero (inference on variable significance) and different from the global trend (inference on local departure). In SVC, it is easy to conduct inference on these aspects by constructing 95% or 99% credible intervals based on the posterior samples. In GWR, although it is possible to construct 95% or 99% confidence intervals, the validation of this depends on the estimation of standard errors. Wheeler and Calder (2007) have noted that GWR tends to under-estimate the standard errors as the coverage probability for the 95% confidence intervals is far below 95% when no exogenous collinearity is imposed.

In SVC, to test whether the local coefficients are significantly different from the average trends, we construct 95% credible intervals for each spatial random effects (w(s)) at each location, we find 20 to 30 sites with significant local departures (the 95% credible intervals do not contain zero) for each w(s). While in GWR, even based on the modified local standard error calculation approach (Fotheringham et al. 2002, p 55), we find more than 100 locations with the 95% confidence intervals not containing the global coefficient means for each predictor variable indicating more than 100 locations with significant local variations. This tendency of over-estimating of spatial heterogeneity for GWR should be considered when drawing conclusions.

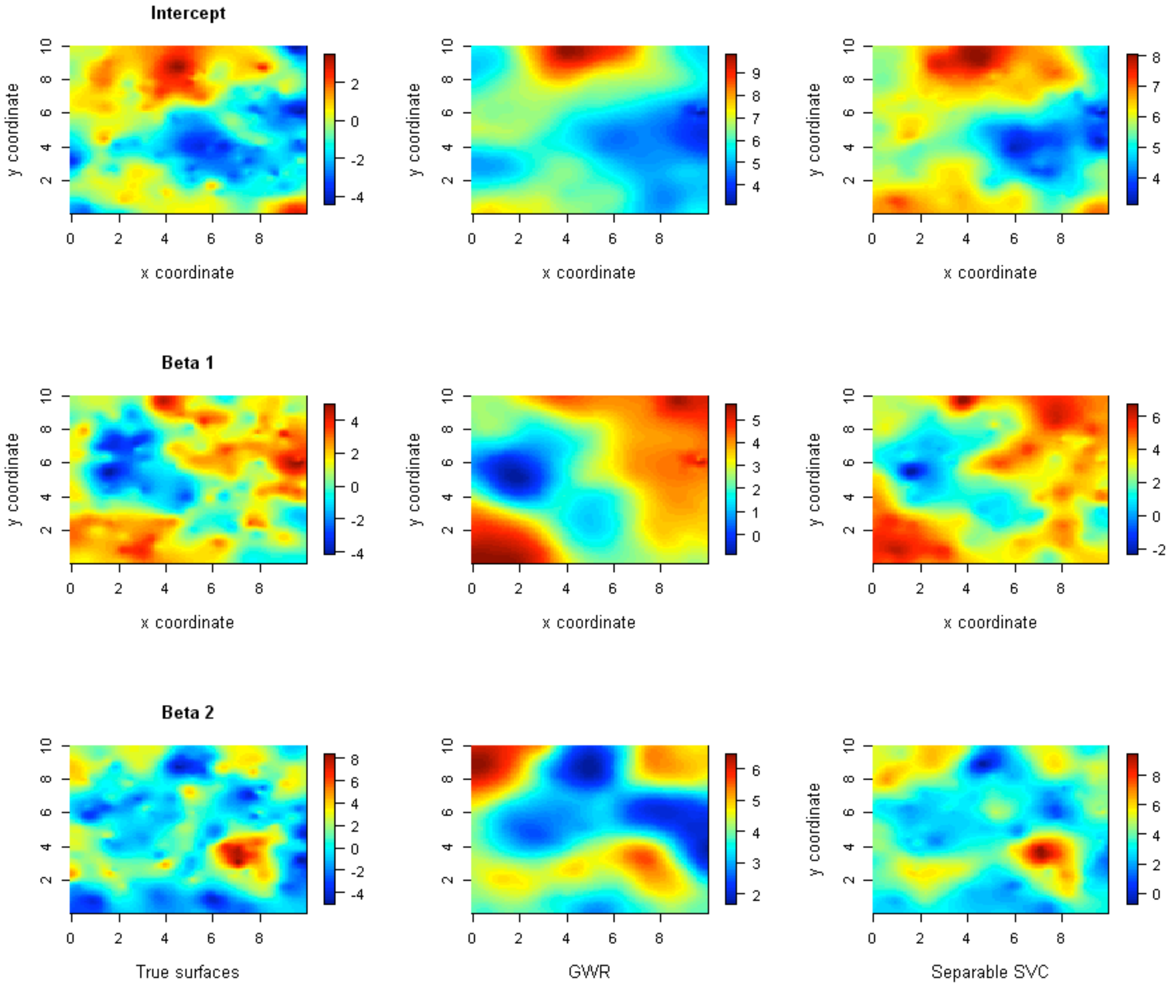

Figure 1. Interpolated surfaces of spatially varying coefficients based on the true values (the first column), estimates from GWR (the second column) and SVC (the last column).

3.2 Quantifying prediction uncertainty in SVC

Similarly, based on the posterior samples, we can easily quantify the prediction uncertainty by constructing 95% or 99% prediction credible intervals for each location. We predict the outcome variable in the handout set using both GWR and SVC. The root mean squared error of prediction (RMSEP) for GWR is 2.48 while RMSEP for SVC is 2.79 indicating GWR is slightly more accurate than SVC.

However, when quantifying the prediction uncertainty, again GWR tends to under-estimate the prediction uncertainty, that is, GWR gives much narrower 95% prediction intervals (for the calculation approach see Harris et al. 2011, p 124) as shown in Figure 2. Another fact is that although GWR seems more accurate than SVC, there are only 53% locations whose true values are within the 95% prediction intervals in GWR, which is much lower than in SVC (about 72%).

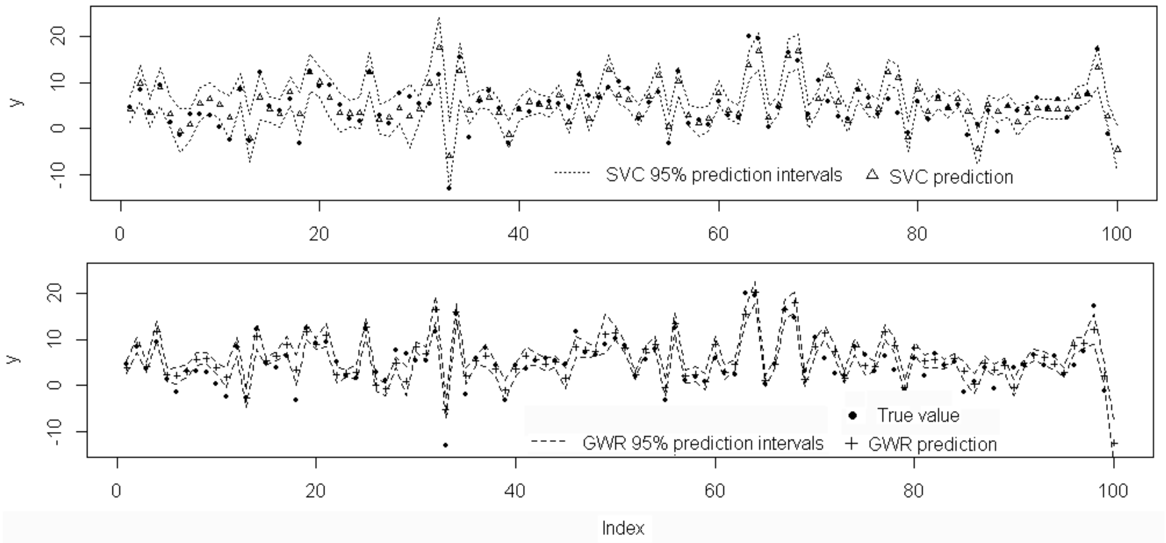

Figure 2. Comparing predictions for the outcome variable and the associated 95% prediction intervals in the handout set between SVC and GWR.

4. Conclusion

In this paper we have briefly described the Bayesian SVC, a global approach to model spatial heterogeneity and illustrated its relative advantages over GWR, a local spatial modelling technique, based on a synthetic data set. The results show that SVC in the Bayesian paradigm is straightforward and reliable to conduct inferences on spatial heterogeneity and to quantify prediction uncertainty by constructing credible intervals based on posterior samples.

References

Finley A O (2011) Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence Methods in *Ecology and Evolution* **2** pp143-154

Fotheringham A S, Brunsdon C and Charlton M (2002) *Geographically weighted regression: the analysis of spatial varying relationships*, John Wiley and Sons, Chichester.

Gelfand A E, Kim H, Sirmans C F and Banerjee S 2003 Spatial modelling with spatially varying coefficient processes *Journal of the American Statistical Association* **98** pp387-396

Harris P, Brunsdon C and Fotheringham A S (2011) Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor *Stochastic Environment Research and Risk Assessment* **25** pp123-138

Nakaya T 2008 Geographically Weighted Regression. In K Kemp (ed.) *Encyclopedia of Geographic Information Science*. Sagem Thousand Oaks, CA, Sage, pp179-184

Plummer M, Best N, Cowles K and Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC *R News* **6** pp7-11

Wheeler D and Calder C (2007) An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems* **9** pp145-166

Wheeler D and Waller L (2009) Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests *Journal of Geographical Systems* **11** pp1-22

Biography

Guanpeng Dong is a second year postgraduate student sponsored by the ESRC within the University of Bristol's Advanced Quantitative Methods training centre, looking at methods to model spatial heterogeneity. His first paper, Using Contextualised Geographically Weighted Regression to Model Spatial Heterogeneity of Land Prices in Beijing, China, is under review.

Rich Harris is a Reader in Quantitative Geography. Recent work has been in geographies of education, focusing on choice and markets in educational systems, measures of segregation, supporting the transition of pupils from primary to secondary schools, and on supporting quantitative and statistical literacy amongst geographers and undergraduate social scientists.