Title
A Closer Examination of Spatial-Filter-Based Local Models

Journal
International Conference on GIScience Short Paper Proceedings, 1(1)

Authors
Oshan, Taylor
Fotheringham, A. Stewart

Publication Date
2016

DOI
10.21433/B31104t0t6ds

Peer reviewed

A Closer Examination of Spatial-Filter-Based Local Models

Taylor Oshan, A. Stewart Fotheringham

Arizona State University, Tempe, Arizona 85281
Email: {toshan; stewart.fotheringham}@asu.edu

Abstract

Local modeling is a spatial analysis technique that explores spatial non-stationarity in data-generating processes. Geographically weighted regression (GWR) is one method that has been widely applied across domains, which has helped uncover local spatial relationships and generally increases model goodness-of-fit. Despite this, some have criticized GWR for being extra susceptible to issues such as multicollinearity and spatial autocorrelation. Spatial-filtering-based local regression (SFLR) has been suggested as a solution. While SFLR is claimed to be approximately equivalent to GWR, it is also touted as superior. Therefore, it is of interest to compare the output from these two techniques. We do this by examining how well both techniques replicate the known coefficient values derived by simulating data that is representative of spatially varying processes. The results indicate that the original SFLR specification is prone to overfitting, while an alternative specification that minimizes the mean square error produces coefficients similar to GWR.

1. Introduction and Background

Non-stationarity in data-generating processes goes largely undetected in traditional global models. Hence, local models, which explicitly allow regression coefficients to vary over space, are necessary to capture such heterogenous processes. One local modeling technique that has become particularly popular is geographically weighted regression (GWR) (Fotheringham *et al.* 2002). Despite its usefulness, GWR has been critiqued as being highly susceptible to multicollinearity (Wheeler & Tiefelsdorf 2005), whereby multicollinearity amongst explanatory variables causes intolerable levels of correlation amongst GWR coefficients. However, results show that when the sample size is large, GWR is robust to even remarkably high levels of multicollinearity (Páez *et al.* 2011; Fotheringham & Oshan Submitted). Still, the work of Wheeler and Tiefelsdorf (Wheeler & Tiefelsdorf 2005) sparked much subsequent critiques of GWR. Specifically, spatial-filter-based local regression (SFLR) has been suggested as a superior alternative to GWR (Griffith 2008). While Griffith (2008) posits that SFLR and GWR are approximately equivalent, he also points out that local coefficients produced from GWR and SFLR are minimally correlated, suggesting that the two models are producing much different results, thereby implying a contradiction within the SFLR method. To the knowledge of the authors, the SFLR method has not been applied outside its original conception (Griffith 2008), while GWR has produced agreeable results in many studies. Therefore, the primarily goal of this paper is to employ simulated data in order to test which of the techniques can more reliably estimate the true coefficients of non-stationary processes. Some modifications of the SFLR routine are investigated and several issues of the SFLR framework are highlighted.

A basic GWR model may be specified as

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where y_i is the dependent variable at location i, β_{i0} is the intercept coefficient at location i, x_{ik} is the kth explanatory variable at location i, β_{ik} is the kth local regression coefficient for the kth explanatory variable at location i, and ε_i is the random error term associated with location i. A kernel function is applied at each location, which is chosen either using cross-validation or maximizing a model fit criterion, in order to weight nearby observations based on their proximity to the calibration location. Consequently, GWR can generate an ensemble of models by creating local subsets of data.

In contrast, SFLR is based on the interpretation that the eigenvectors of a modified connectivity matrix are the set of possible orthogonal and uncorrelated map patterns (i.e., degree of spatial autocorrelation) (Griffith 1996). Griffith proposes that interaction terms between each explanatory variable and each eigenvector can be utilized to create local coefficients that are approximately equal to those from GWR. Therefore, the following specification is derived

$$y_i = \beta_0 1 + \sum_{p=1}^{P} X_p \cdot 1\beta_p + \sum_{k=1}^{K} E_k \beta_{E_k} + \sum_{p=1}^{P}\sum_{k=1}^{K} X_p \cdot E_k \beta_{pE_k} + \varepsilon \tag{2}$$

where $\beta_0 1$ is the intercept, $\sum_{p=1}^{P} X_p \cdot 1\beta_p$ represents the explanatory variables and their coefficients, $\sum_{k=1}^{K} E_k \beta_{E_k}$ denotes a subset of the possible eigenvectors and their coefficients and $\sum_{p=1}^{P}\sum_{k=1}^{K} X_p \cdot E_k \beta_{pE_k}$ are a subset of the possible interaction terms between explanatory variables and eigenvectors and their coefficients. An important step in the spatial-filter-based framework is selecting a subset of the eigenvectors and interaction terms. The methodology entails a forward stepwise regression variable selection algorithm amongst all interaction terms and eigenvectors based on a statistical significance criterion. For each iteration of the routine, each candidate variable is tested in the model and the one that produces the smallest p-value is selected. At the end of each iteration, any variable that produces a p-value greater than the threshold of 0.1 is removed. The algorithm continues until no variable can be added that produces a p-value less than the threshold, at which point the routine terminates and the local coefficients can be processed.

2. Methods and Results

A dataset with known properties was derived by simulating 2,500 observations for the cells of a 50 by 50 grid. The observations within the cells were generated using the following equation

$$Y = B_{0i} + B_{1i}X_{1i} + B_{2i}X_{2i} + \varepsilon_i \tag{3}$$

where Y are the generated observation, B_0, B_1, and B_2 are *known* locally varying coefficients, X_1 and X_2 are variables drawn form random normal distributions, ε is a random normal error term, and i is the index for each of the 2,500 locations. The spatial distributions of B_0, B_1, and B_2 llustrated in figure 1A display varying levels of non-stationarity, which were created by drawing random values from normal distributions that vary over space and, therefore, are not a result of either the GWR or SFLR specifications. Since these surfaces are based on local processes, they result in MC values that signify very strong spatial autocorrelation (0.982, .947, and .912 for B_0, B_1, and B_2, respectively).

Griffith (2008) previously insinuated that the SFLR framework was superior to GWR based on the fact that the SFLR estimated coefficients "remain unbiased, yield a better global model fit, are polluted with considerably lower levels of spatial autocorrelation, and, for the most part, display little relationship to the GWR coefficients". Therefore, SFLR and GWR were both estimated using the simulated data to compare the resulting coefficient surfaces which can be seen in figures 1B and 1C. It was noticed in figure 2 that the mean square error (MSE)

Table 1: Comparison of MSE, MC, and model fit for models.

	MSE b_0	MSE b_1	MSE b_2	MC b_0	MC b_1	MC b_2	adj. R^2	AIC
GWR	3.239	0.00129	0.00401	0.979	0.994	0.994	0.888	21164.561
SFLR	72.725	0.0130	0.0123	0.651	0.809	0.905	0.946	20116.961
SFLR*	7.498	0.00170	0.00380	1.0132	1.0097	1.00210	0.906	21250.258

of the SFLR coefficients and the known coefficients decreased after about the first 30 terms were selected. However, even as model fit increased, additional model terms increased the MSE, indicating overfitting. Therefore, a new MSE-minimizing stepwise selection criterion (SFLR*) was employed that sought to terminate the algorithm before overfitting occurred. The resulting coefficients are displayed in figure 1D.

Overall, the model fit, coefficient accuracy, and spatial autocorrelation statistics provided in table 1 indicate that SFLR is not approximately equivalent to GWR and that GWR produces less error in estimated coefficients compared to SFLR. In contrast, the SFLR* model produces extremely similar model fit, coefficient estimates, and spatial autocorrelation statistics compared to GWR, though there is no indication of superiority. In fact, model fit statistics that account for model complexity, such as the AIC values presented in table 1 tend to favor GWR (i.e., lower values), since it utilizes far fewer covariates.

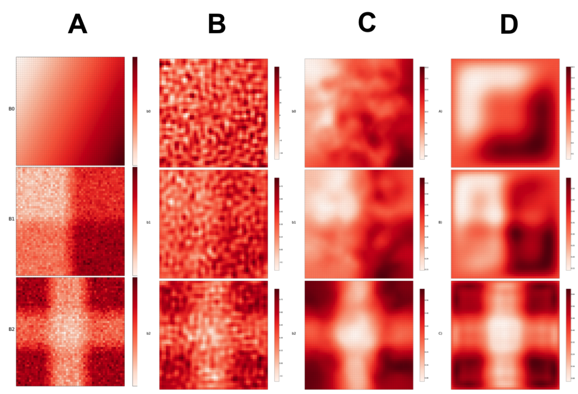

Figure 1: Coefficient surfaces for A) simulated data B) SFLR estimates; C) GWR estimates; D)SFLR* estimates. Top row corresponds to B_0, middle row to B_1 and bottom row to B_2. Low coefficient values are shaded lighter while high values are shaded darker.

3. Discussion & Conclusion

In its original conception, SFLR does not seem to be a competitor of GWR. Using Griffith's (2008) stepwise selection routine produces overfitted models where a severe loss of coefficient accuracy occurs. Griffith was correct in that the SFLR framework and GWR do not produce similar results, though this is only due to an inherent flaw in his methodology that

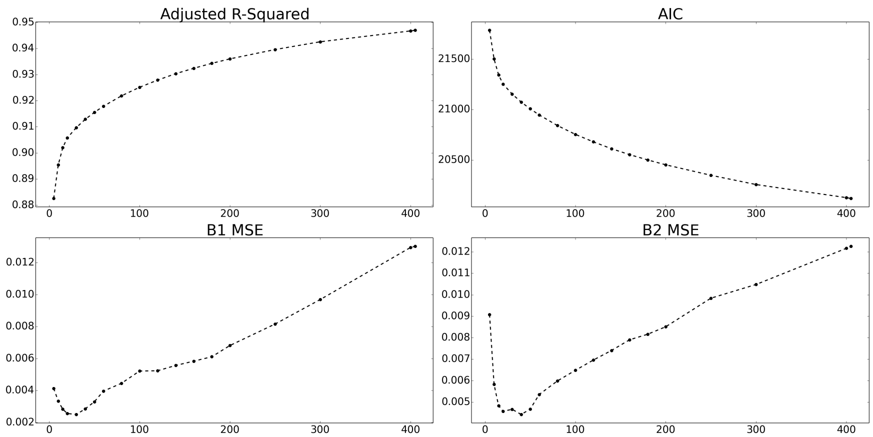

Figure 2: Adjusted R-squared, AIC, and MSE for SFLR model with increasing stepwise selection iterations.

causes overfitting, and can require lengthy compute times of up to several days on a standard notebook computer. It turns out that the SFLR framework *can* be approximately equivalent to GWR when a MSE-minimizing stepwise selection routine (SFLR*) is employed. By doing so, the SFLR* framework produces coefficient estimates that are similar to those from GWR in magnitude, overall accuracy, spatial autocorrelation, and that yield a similar model fit. The computation time is also drastically reduced from days to only minutes, which is comparable to GWR. It seems then that the SFLR* framework provides a sort of discrete spatial weighting mechanism in the form of a subset of eigenvectors and interaction terms, somewhat akin to GWR's continuously defined kernel function weighting mechanism. Given the potential of both GWR and SFLR* to produce such similar results, this provides strong evidence that there is no *a priori* disadvantage to local coefficients displaying strong spatial autocorrelation. Despite the promising results from the SFLR* specification, it is uncertain how to deploy the method when there are no known coefficients because their MSE cannot be computed. Other drawbacks to the SFLR framework include the lack of a means of testing for statistical significance of the estimated coefficients, potential issues of replicability of the methodology, and uncertainty of its robustness to different types of spatial weight matrices.

References

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Fotheringham, A. S., & Oshan, T. (Submitted). Geographically weighted regression and multiocollinearity: dispelling the myth. *Journal of Geograpical Systems*.

Griffith, D. A. (1996). Spatial Autocorrelation and Eigenfunctions Of The Geographic Weights Matrix Accompanying Geo-Referenced Data. *Canadian Geographer / Le GÃl’ographe canadien*, *40*(4), 351–367.

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

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., & Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, *7*(2), 161–187.