Calibrating a Geographically Weighted Regression Model with Parameter-specific Distance Metrics

Article · December 2015

DOI: 10.1016/j.proenv.2015.05.011

Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 110 – 115

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Calibrating a Geographically Weighted Regression Model with Parameter-Specific Distance Metrics

Binbin Lu[a]*, Paul Harris[b], Martin Charlton[c], Chris Brunsdon[c]

[a]*School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.*
[b]*Rothamsted Research, North Wyke, Okehampton, Devon, UK*
[c]*National Centre for Geocomputation, Maynooth University, Maynooth, Co. Kildare, Ireland*

Abstract

Geographically Weighted Regression (GWR) is a local technique that models spatially varying relationships, where Euclidean distance is traditionally used as default in its calibration. However, empirical work has shown that the use of non-Euclidean distance metrics in GWR can improve model performance, at least in terms of predictive fit. Furthermore, the relationships between the dependent and each independent variable may have their own distinctive response to the weighting computation, which is reflected by the choice of distance metric. Thus, we propose a back-fitting approach to calibrate a GWR model with parameter-specific distance metrics. To objectively evaluate this new approach, a simple simulation experiment is carried out that not only enables an assessment of prediction accuracy, but also parameter accuracy. The results show that the approach can provide both more accurate predictions and parameter estimates, than that found with standard GWR. Accurate localised parameter estimation is crucial to GWR's main use as a method to detect and assess relationship non-stationarity.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: Non-stationarity, GWR, Parameter-Specific Distance Metrics, Simulation Experiment,

1. Introduction

A number of localized regression techniques have been proposed to account for spatial non-stationarity or spatial heterogeneity in data relationships, one of which is geographically weighted regression (GWR) [1]. Key to GWR is a 'bump of influence' around each local regression point: where nearer observations have more influence in estimating

* Binbin Lu. Tel.: +86-27-68770771; fax: +86-27-68778086.
E-mail address: binbinlu@whu.edu.cn

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.011

the local set of parameters than do observations farther away [2]. This is described by a kernel weighting function based on distances between model calibration points and observation points. Euclidean distance (ED) is traditionally used as default in calibrating a GWR model. However, empirical work has shown that the use of non-Euclidean distance metrics (like network distance and travel time metrics) in GWR can improve model fit [3, 4]. Furthermore, the relationship between the dependent and each independent variable may have its own distinctive response to the weighting computation.

Some related and important studies have been done in this respect, where the bandwidth of the kernel function is allowed to vary across relationships. Brunsdon et al. [5] introduced mixed GWR, which considers some data relationships as global (or fixed), and the rest as local (but each at the same spatial scale). Yang [6] generalizes the mixed GWR model by allowing each data relationship to operate at its own (and commonly different) spatial scale. In this study, we enhance both studies, where the choice of distance metric is also allowed to vary over different parameter estimates in the same model. We hypothesize that each independent/dependent variable pair in the GWR model may correspond to different "optimal" distance metrics, and then calibrate GWR with parameter-specific distance metrics (PSDM-GWR). A back-fitting approach inherited from mixed GWR is adjusted for the PSDM-GWR model calibration. PDSM-GWR is evaluated via a simple simulation experiment. All of the modelling functions used in this article can be found in the **GWmodel** package [7, 8] in **R** [9], which is an integrated framework for handling spatially-varying structures, via a wide range of geographically weighted models.

2. Methodology

GWR estimates a localized set of regression parameters in order to assess the possibility of spatially-varying relationships. The basic formula of a GWR model can be written as:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \tag{1}$$

where y_i is the dependent variable at location *i*, x_{ik} is the value of the *k*th explanatory variable at location *i*, β_{i0} is the intercept parameter at location i, β_{ik} is the local regression parameter (or coefficient) for the *k*th explanatory variable at location *i*, and ε_i is the random error at location *i*. At each location, the model is calibrated by a weighted least squares approach, of which the matrix expression is:

$$\hat{\boldsymbol{\beta}}_{\mathbf{i}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}_{\mathbf{i}}\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}_{\mathbf{i}}\mathbf{y} \tag{2}$$

where $\mathbf{W_i}$ is the diagonal matrix denoting the geographical weightings for each observation data (sub-)set for regression point *i*. In a standard GWR calibration, $\mathbf{W_i}$ is calculated via a kernel function whose bandwidth, is customarily selected via a leave-one-out cross-validation (CV) approach [10] or an Akaike Information Criterion (AIC) approach [11].

For this study, the GWR technique is extended to PSDM-GWR, where the back-fitting algorithm used in mixed GWR [5] and (similarly) in flexible bandwidth GWR [6] is adjusted for PSDM-GWR calibration. If we assume that the specific distance metrics are respectively, $DM_0, DM_1, \cdots, DM_m$ for estimating their corresponding parameters, and the hat matrix for each parameter estimates is defined as $\boldsymbol{S}_0, \boldsymbol{S}_1, \cdots, \boldsymbol{S}_m$, then eq.(1) can be re-written as:

$$\hat{\mathbf{y}} = \sum_{j=0}^{m} \hat{\mathbf{y}}_j = \sum_{j=0}^{m} S_j \mathbf{y} \tag{3}$$

Then the back-fitting procedure to calibrate PSDM-GWR can be carried out in the following steps:

Step 1. Initialize values of $\hat{\boldsymbol{y}}_0, \cdots, \hat{\boldsymbol{y}}_m$, with $\hat{\boldsymbol{y}}_0^{(0)}, \cdots, \hat{\boldsymbol{y}}_m^{(0)}$;

Step 2. Set *i=1*;

Step 3. Calculate $\widehat{\boldsymbol{y}}_j^{(i)} = \boldsymbol{S}_{\mathrm{j}}\left[\mathbf{y} - \sum_{\mathrm{k}\neq\mathrm{j}} Latestyhat\left(\widehat{\boldsymbol{y}}_k^{(i-1)}, \widehat{\boldsymbol{y}}_k^{(i)}\right)\right]$, where the $Latestyhat$ function is defined in eq.(4), and $\boldsymbol{S}_{\mathrm{j}}$ is calculated using DM_j and a given bandwidth bw_j;

$$Latestyhat\left(\widehat{\mathbf{y}}_k^{(i-1)}, \widehat{\mathbf{y}}_k^{(i)}\right) = \begin{cases} \widehat{\mathbf{y}}_k^{(i)}, \text{ if } \widehat{\mathbf{y}}_k^{(i)} \text{ exists} \\ \widehat{\mathbf{y}}_k^{(i-1)}, \text{otherwise} \end{cases} \tag{4}$$

Step 4. Repeat Step 3 from 0 to m;
Step 5. Calculate the residual sum of squares $RSS^{(i)}$ between $\mathbf{y}$ and $\widehat{\mathbf{y}}^{(i)}$, and set $i=i+1$;
Step 6. Return to Step 3 unless $RSS^{(i)}$ converges to $RSS^{(i-1)}$.
In this procedure, the choice of initial guesses is open. Here we use the results form a standard GWR calibration (eq.(2)) as starting values in Step 1. The sensitivity of the back-fitting algorithm to different initial guesses is currently under consideration, but poor initial guesses will undoubtedly affect the speed of convergence.

3. Case study with simulated data

As an introductory assessment of the PSDM-GWR model, we use simulated data. For this basic simulation experiment, a point data set of size 25*25 is generated on a square grid, of which the coordinates in two dimensions range from 10 to 100. For each cell, two predictor variables x_1 and x_2 are independently drawn from a uniform distribution as a random numeric vector ranging from 1 to 100, as shown in Fig. 1.

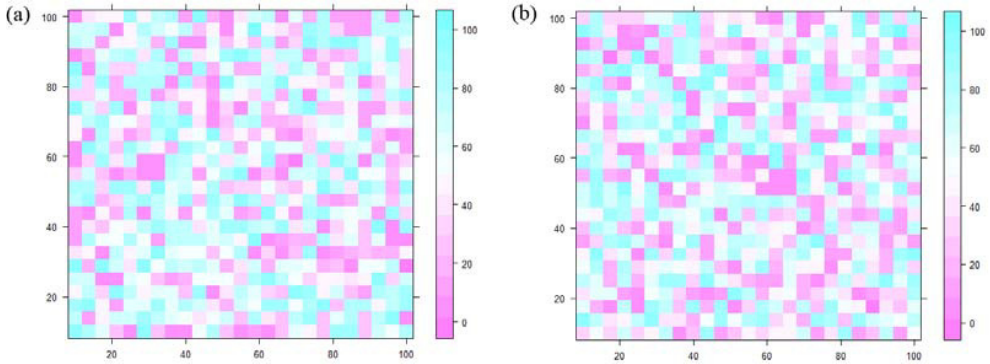

Fig. 1 (a) Surface for the random predictor x_1; (b) Surface for the random predictor x_2.

The process to generate each realisation of this simulation experiment is defined as follows:

$$y = \beta_1 x_1 + \beta_2 x_2 \tag{5}$$

$$\beta_1 = 2, \beta_2 = \log(u+v) \tag{6}$$

where the dependent variable y is naturally generated from eq. (5), which itself consists of a stationary (single) parameter β_1 and a non-stationary parameter β_2 , as found from the equations in (6). It is a fairly simple case study, but represents clearly different varying relationships between y and x_1 and between y and x_2. Observe that we do not simulate an intercept parameter, β_0. The corresponding surfaces of β_2 and y are visualized in Fig. 2.

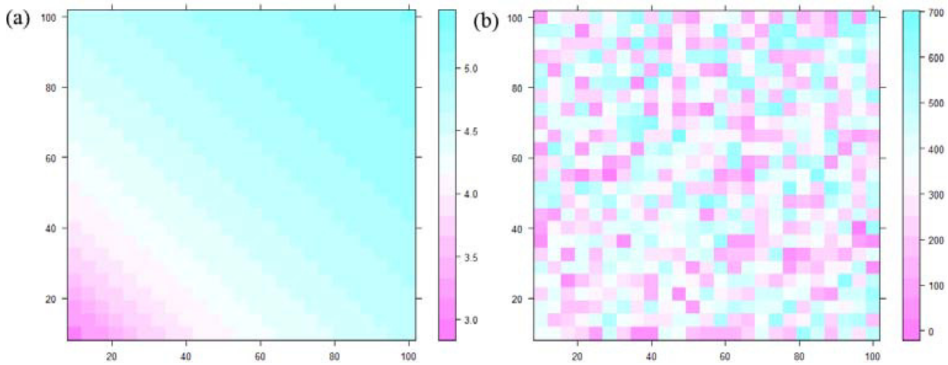

Fig. 2. (a) Surface for the coefficient β_2; (b) Surface for the dependent variable *y*.

Using one realisation of the simulation, we calibrate the model shown in eq. (5) via both standard GWR and PSDM-GWR. For standard GWR, ED is used to estimate both β_1 and β_2; which is the standard approach. However for PSDM-GWR, we use a zero distance matrix (i.e. assuming the distance between any pair of points is zero, i.e. a simple non-ED metric) to estimate β_1 and a ED matrix to estimate β_2. Thus it represents a simple form of PSDM-GWR and is chosen to demonstrate its potential. For an objective comparison, we use the same fixed bandwidth for both GWR calibrations, which is selected by an AIC approach using the standard GWR model.

The results are presented in Table 1, where a reduction in RSS indicates that PSDM-GWR provides more accurate predictions than standard GWR. Fig. 3 plots the estimated parameters β_1 and β_2 from both calibrations. As would be expected, PSDM-GWR provides a highly accurate estimate of the stationary (constant) parameter β_1, with $\hat{\beta}_1 = 1.998$; whilst similarly as expected, standard GWR provides a non-constant estimation of β_1 and as such, is relatively inaccurate. In terms of β_2, both models provide similar estimates, but the estimates from PSDM-GWR appear slightly closer to the real values than that found with standard GWR. Tentatively, this simple experiment suggests that PSDM-GWR can also provide more accurate parameter estimates than that found with standard GWR.

Table 1. Model calibrations via standard GWR and PSDM-GWR

	Distance metric(s)	Kernel function	Bandwidth	RSS
Standard GWR	ED for estimating both β_1 and β_2	Gaussian function with a fixed bandwidth selected by AICc approach in a standard way	3.54	446.11
PSDM-GWR	Zero distance matrix for estimating β_1, ED matrix for estimating β_2			418.20

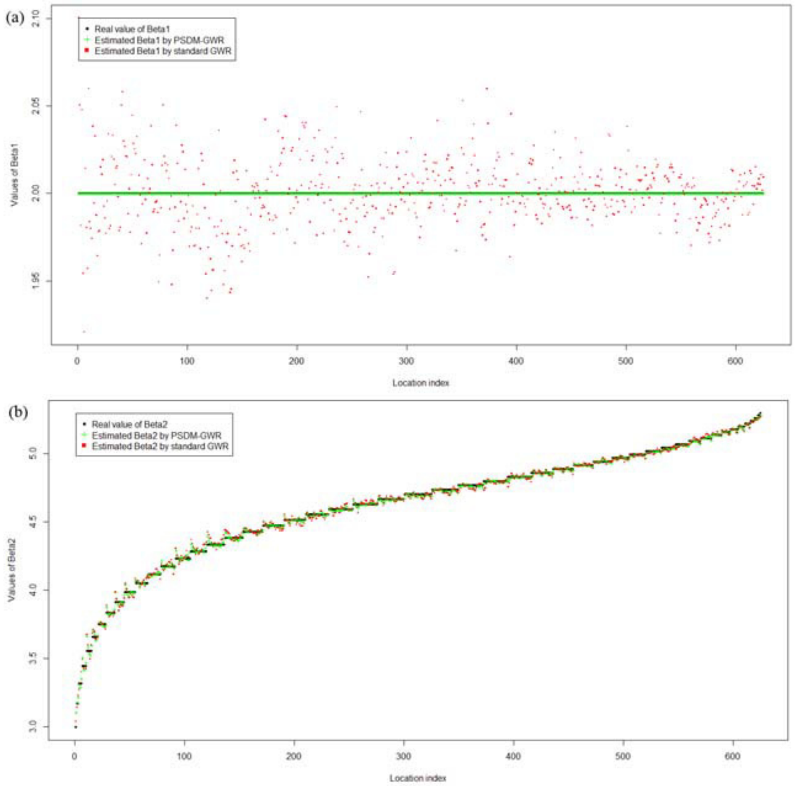

Fig. 3. (a) Real values of β_1 and estimations from standard GWR and PSDM-GWR; (b) Real values of β_2 and estimations from standard GWR and PSDM-GWR.

4. Concluding remarks

In this study, we proposed a back-fitting algorithm for PSDM-GWR. Via a simulation study, we have shown that PSDM-GWR can provide more accurate predictions and parameter estimates than standard GWR. However, this can only be considered as preliminary findings, as:

- The form of the PSDM-GWR model used in this study is just a specific case of a mixed GWR model. In this respect, a more involved simulation study is required using (novel) PSDM-GWR specifications that do not mimic existing GWR constructions.
- The way to define or select a distance metric for an independent variable within a given PSDM-GWR model is key and requires refinement.
- PSDM-GWR also needs to demonstrate its practical worth within an empirical case study.
- The approach could be meshed with that of Yang [6], where bandwidths vary across relationships.

Acknowledgements

Research presented in this paper is funded by National Natural Science Foundation of China (NSFC: 41401455). The authors gratefully acknowledge this support.

References

[1].Brunsdon, C., A.S. Fotheringham, and M.E. Charlton, *Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity.*

Geographical Analysis, 1996. **28**(4): p. 281-298.

[2].Fotheringham, A.S., M.E. Charlton, and C. Brunsdon, *Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis.* Environment and Planning A, 1998. **30**(11): p. 1905-1927.

[3].Lu, B., M. Charlton, and A.S. Fotheringham, *Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data.* Procedia Environmental Sciences, 2011. **7**(0): p. 92-97.

[4].Lu, B., et al., *Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data.* International Journal of Geographical Information Science, 2014. **28**(4): p. 660-681.

[5].Brunsdon, C., A.S. Fotheringham, and M. Charlton, *Some Notes on Parametric Significance Tests for Geographically Weighted Regression.* Journal of Regional Science, 1999. **39**(3): p. 497-524.

[6].Yang, W., *An Extension of Geographically Weighted Regression with Flexible Bandwidths*, in *Centre for GeoInformatics*. 2014, University of St Andrews: St Andrews, UK.

[7].Gollini, I., et al., *GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models.* Journal of Statistical Software, 2015. **63**(17): p. 1-50.

[8].Lu, B., et al., *The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models.* Geo-spatial Information Science, 2014. **17**(2): p. 85-101.

[9].R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2013, R Foundation for Statistical Computing: Vienna, Austria.

[10].Farber, S. and A. Páez, *A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations.* Journal of Geographical Systems, 2007. **9**(4): p. 371-396-396.

[11].Fotheringham, A.S., C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: the analysis of spatially varying relationships*. 2002, Chichester: Wiley.