Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

# Improvements to the calibration of a geographically weighted regression with parameter-specific distance metrics and bandwidths

Binbin Lu[a,b,*], Wenbai Yang[c], Yong Ge[d], Paul Harris[e]

[a] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[b] Key Laboratory for National Geographic Census and Monitoring, National Administration of Surveying, Mapping and Geoinformation, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[c] Reed Elsevier Information Technology (Beijing) Co., Ltd., Beijing, China
[d] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[e] Sustainable Soils and Grassland Systems, Rothamsted Research, North Wyke, Okehampton, UK

ARTICLE INFO

*Keywords:*
Local regression
Spatial heterogeneity
Bandwidth selection
Multi-scale
GWmodel

ABSTRACT

In standard geographically weighted regression (GWR), the spatially-varying relationships between the dependent and each independent variable are explored under a constant and fixed scale, but for many processes their variation intensity may differ with respect to location and direction. To address this short-coming, a GWR model with parameter-specific distance metrics (PSDM GWR) can be used, which by default, also specifies parameter specific bandwidths. In doing so, PSDM GWR provides a scale-dependent extension of GWR. Commonly however, an ideal distance metric for a given independent variable parameter is not immediately obvious. Thus, in this article, PSDM GWR is investigated with respect to distance metric choice. Here, it is demonstrated that the optimum (distance metric specific) bandwidth corresponding to a given independent variable remains essentially constant, independent of the choices made for the other independent variables. This result allows for a considerable saving in computational overheads, permitting a much simpler searching procedure for multiple bandwidth optimization. Results are first demonstrated empirically, and then a simulation experiment is conducted to objectively verify the same findings. Computational savings are vital to the uptake of PSDM GWR, where ultimately, it should be considered the default choice in any GWR-based study of spatially-varying relationships, as standard GWR, mixed (or semi-parametric) GWR, flexible bandwidth (or multi-scale) GWR and the global regression are specific cases thereof.

## 1. Introduction

As indicated by Goodchild (2004), spatial heterogeneity in geographic variables or relationships, as a corollary of uncontrolled variation, requires a spatially-bounded analysis, “move the study area, and the results will change”. In this context, numerous localized methods have been proposed that produce spatially-varying outputs instead of a ‘one-size-fits-all’ output from a global, often non-spatial, method (Fotheringham & Brunsdon, 1999). Early spatially-localized techniques include the expansion method (Casetti, 1972), kriging with local variograms (Haas, 1990, 1996), a local multilevel model (Jones, 1991), local indicators of spatial association (Anselin, 1995; Getis & Ord, 1992), a spatially-adaptive filtering model (Gorr & Olligschlaeger, 1994), geographically weighted (GW) regression (GWR) (Brunsdon, Fotheringham, & Charlton, 1996; McMillen, 1996), GW summary statistics (Brunsdon, Fotheringham, & Charlton, 2002), GW principal components analysis (Fotheringham, Brunsdon, & Charlton, 2002; Harris, Brunsdon, & Charlton, 2011) and Bayesian space-varying coefficient (SVC) models (Assunção, 2003; Gelfand, Kim, Sirmans, & Banerjee, 2003). More recent techniques include GW discriminant analysis (Brunsdon, Fotheringham, & Charlton, 2007), SVC eigenvector spatial filtering (Griffith, 2008; Murakami, Yoshida, Seya, Griffith, & Yamagata, 2017) and GW quantile regression (Chen, Deng, Yang, & Matthews, 2012). Many of the GW-based models have been integrated into the **GWmodel R package** (Gollini, Lu, Charlton, Brunsdon, & Harris, 2015; Lu, Harris, Charlton, & Brunsdon, 2014).

The focus of this study lies with GWR, which is an increasingly popular technique for modeling heterogeneous processes across many research domains (Fotheringham, Crespo, & Yao, 2015). It allows for the investigation of spatially-varying relationships via the fitting of

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.
*E-mail address:* binbinlu@whu.edu.cn (B. Lu).

https://doi.org/10.1016/j.compenvurbsys.2018.03.012
Received 11 July 2017; Received in revised form 13 March 2018; Accepted 31 March 2018

individual localized linear regressions at focal locations. Here a 'bump of influence' allows nearer observations to have a greater influence in the local regression fit than observations further away (Fotheringham et al., 2002). This is achieved by weighting observations via a kernel weighting scheme, where the kernel can be any distance-decay function, which is non-increasing, real, and bounded from 0 to 1 (Cho, Lambert, & Chen, 2010). Crucial to GWR is the choice of distance metric and the size of the bandwidth. The latter describes the rate of distance decay for the data weightings: a large bandwidth tends to smooth out variation in the local regression parameter estimates (larger bias), while a small bandwidth tends to sharpen them (larger variance).

For distance metrics in GWR, Lu, Charlton, Brunsdon, and Harris (2016), Lu, Charlton, and Fotheringham (2011), Lu, Charlton, Harris, and Fotheringham (2014), Lu, Harris, et al. (2014) proposed the use of non-Euclidean distance (non-ED) metrics, and found model fit to be significantly improved to that found with the Euclidean distance (ED); and intriguingly, Huang, Wu, and Barry (2010) and Fotheringham et al. (2015) in space-time extensions of GWR, defined the temporal metric using distances rather than time. For bandwidths, a fixed distance or a fixed number of nearest neighbors (adaptive) is specified (Fotheringham et al., 2002; Wheeler & Páez, 2010). An optimum bandwidth is then determined by minimizing a model fit statistic (Loader, 1999), such as a leave-one-out cross-validation (CV) score (Brunsdon et al., 1996; Cleveland, 1979) or the Akaike Information Criterion (AIC) (Akaike, 1973; Fotheringham et al., 2002). Farber and Páez (2007) proposed two modified CV approaches to relieve excessive influence of the CV score on any single (anomalous) focal point. Cho et al. (2010) optimized the bandwidth via minimizing a spatial error Lagrange Multiplier, and so reduce error autocorrelation in the resultant GWR model.

However, all such GWR calibrations naively assume a uniform scale or magnitude of the non-stationarities for all dependent-independent relationships. It is more likely that the variation intensity of these relationships is different. That is, relationships are not only non-stationary, but operate at different spatial scales. In this respect, Brunsdon, Fotheringham, and Charlton (1999) introduced mixed GWR that treats some data relationships as global (or fixed), and the rest as local (but each at the same spatial scale). Yang (2014) extended this notion via flexible bandwidth GWR (FBGWR), also known as multi-scale GWR (Fotheringham, Yang, & Kang, 2017), where each dependent-independent relationship operates at its own (and commonly different) spatial scale, via specifying a different bandwidth for each independent variable. Lu, Brunsdon, Charlton, and Harris (2017); Lu, Harris, Charlton, and Brunsdon (2015) combined FBGWR with the use of different distance metrics for each relationship, to form the parameter-specific distance metric model (PSDM GWR) of this study.

Unfortunately, a suitable distance metric for a given independent variable is rarely known with clarity, and how to specify a variety of metrics within the same model is difficult, especially when the number of independent variables is large. Distance metrics can vary greatly due to the diversity of the sample data and the complexity of the underlying geography. If metrics are known, they are not always correctly calculated. Although Lu et al. (2017, 2015) introduced PSDM GWR, its computational burden clearly required attention, and as such, only a rather limited form of PSDM GWR was demonstrated. This study now revisits the PSDM GWR model where the selection procedure (search strategy) for the parameter-specific distance metrics is revised and improved. By default, FBGWR is also revisited (where it only specifies ED metrics) and its fitting procedure is similarly improved.

This article is organized as follows. Firstly, we describe PSDM GWR and present a brute-force search strategy. Secondly, we empirically assess all possible distance metric combinations for a PSDM GWR model with ED and travel time (TT) as candidate metrics, and in doing so, we propose a revised search strategy that reduces computational overheads. We also experiment with the Minkowski approach for the same purpose. Thirdly, we objectively endorse the empirical findings via a simulation experiment. Finally, we summarize the study results and the advances therein.

## 2. Methodology

### 2.1. GWR

A standard GWR model can expressed as:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{1}$$

where $y_i$ and $x_{ik}(k = 1, \cdots, m)$are the observations of the dependent variable and the independent variables, respectively at location $i$, $\beta_{ik}(k = 0, 1, \cdots, m)$ is the set of regression parameters estimated at location $i$; and $\varepsilon_i$ is the random error term. It is calibrated by a weighted least squares approach at each regression point, of which the matrix expression is:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_i\mathbf{y} \tag{2}$$

where $\mathbf{X}$ is the matrix of the independent variables with $m + 1$ columns and a column of $1$s for the intercept (if there is one); $\mathbf{y}$ is the vector for the dependent variable; $\mathbf{W}_i$ is a diagonal matrix denoting the geographical weightings for each observation data (sub-)set at regression location $i$, and can be found via a kernel function that generates distance-decay values ranging from 0 to 1. Kernels are commonly specified as Gaussian, exponential, box-car, bi-square and tri-cube (Gollini et al., 2015).

### 2.2. GWR with parameter-specific distance metrics

In standard GWR, the weighting matrix is calculated with a ED metric together with a unique bandwidth. This assumes that 'as the crow-flies' distances are appropriate throughout, and that any dependent/independent variable relationship varies at the same spatial scale. However, the spatially-varying scale or intensity of the different dependent/independent variable relationships may differ, and as such, each should have their own distinct weighting scheme within the same model (Lu et al., 2015; Yang, Fotheringham, & Harris, 2011). Such situations are catered for with PSDM GWR, which is implemented via an adjusted back-fitting algorithm as used in mixed GWR and FBGWR, which are each a particular case of PSDM GWR. The PSDM GWR model can be expressed as:

$$y_i = \beta_{0i}^{(DM_0, bw_0)} + \sum_{k=1}^{m} \beta_{ki}^{(DM_k, bw_k)} x_{ik} + \varepsilon_i \tag{3}$$

where $DM_k$and $bw_k$ $(k = 0, 1, \cdots, m)$ represent the specific distance metric and bandwidth for each independent variable (and intercept) parameter estimate, respectively. A full account of PSDM GWR is provided in Lu et al. (2017).

To choose an optimum bandwidth for each parameter of PSDM GWR, an optimization can be conducted by minimizing the CV score or the corrected AIC (AICc) within the back-fitting iterations (Fotheringham et al., 2017; Lu et al., 2017). Note AICc takes into account the effective sample size. However as spatial autocorrelation is likely, the effective sample size is expected to be much smaller than the nominal sample size. Lu et al. (2017) indicated that the bandwidth for each parameter of a PSDM GWR model would converge quickly to an optimum, regardless of whether the optimization procedure is exhaustive or not. The same can be found in the presentation of FBGWR (Fotheringham et al., 2017). Thus, Lu et al. (2017) proposed a threshold value δ to diagnose bandwidth convergence, i.e. the bandwidth value for a given parameter will be set as fixed, instead of being exhaustively searched for, provided the change of the optimized values in the respective iterations is less than δ. This approach provides some computational saving, but the back-fitting procedure itself still presents a