A Hierarchical and Geographically Weighted Regression Model and Its Backfitting Maximum Likelihood Estimator

Yigong Hu[1]
School of Geographical Sciences, University of Bristol, UK

Richard Harris
School of Geographical Sciences, University of Bristol, UK

Richard Timmerman
School of Geographical Sciences, University of Bristol, UK

Binbin Lu
School of Remote Sensing and Information Engineering, Wuhan University, Hubei, China

Abstract

Spatial heterogeneity is a typical and common form of spatial effect. Geographically weighted regression (GWR) and its extensions are important local modeling techniques for exploring spatial heterogeneity. However, when dealing with spatial data sampled at a micro-level but the geographical locations of them are only known at a higher level, GWR-based models encounter several problems, such as difficulty in establishing the bandwidth. Because data with this characteristic exhibit spatial hierarchical structures, such data can be suitably handled using hierarchical linear modeling (HLM). This model calibrates random effects for sample-level variables in each group to address spatial heterogeneity. However, it does not work when exploring spatial heterogeneity in some group-level variables when there is insufficient variance in each group. In this study, we therefore propose a hierarchical and geographically weighted regression (HGWR) model, together with a back-fitting maximum likelihood estimator, that can be applied to examine spatial heterogeneity in the regression relationships of data where observations nest into high-order groupings and share the same or very close coordinates within those groups. The HGWR model divides coefficients into three types: local fixed effects, global fixed effects, and random effects. Results of a simulation experiment show that HGWR distinguishes local fixed effects from others and also global effects from random effects. Spatial heterogeneity is reflected in the estimates of local fixed effects, along with the spatial hierarchical structure. Compared with GWR and HLM, HGWR produces estimates with the lowest deviations of coefficient estimates. Thus, the ability of HGWR to tackle both spatial and group-level heterogeneity simultaneously suggests its potential as a promising data modeling tool for handling the increasingly common occurrence where data, in secure settings for example, remove the specific geographic identifiers of individuals and release their locations only at a group level.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial modelling, hierarchical data, spatial heterogeneity, geographically weighted regression

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.39

Category Short Paper

Supplementary Material *Software (Source code)*: `https://github.com/hpdell/hgwr`
archived at `swh:1:dir:c9c2bab2a6428b8d3b6d25a3da472653018a7fae`
Text (Blog post): `https://hpdell.github.io/GIScience-Materials/posts/HGWR/`

Funding *Yigong Hu*: Yigong Hu was sponsored by the China Scholarship Council with the University of Bristol (No. 202106270029).

[1] Corresponding author.

1 Introduction

In statistics and data analysis, regression models are powerful tools in examining relationships in data. However, the ordinary linear regression, as a model of global relationships, holds many limitations in dealing with spatial data [5] because the relationship between variables may not keep constant across the whole area. In spatial statistics, this phenomenon is called "spatial heterogeneity" [2]. To uncover such an effect, many local-form spatial statistic methods are proposed to discover underlying spatial heterogeneity in data [5]. The geographically weighted regression (GWR) [3] model and its extensions are popular ones. These methods calibrate a unique model at each location to produce spatially varying coefficients by borrowing samples from its geographical neighbors defined by spatial distances. Shorter distance gives rise to higher weighting. Among its extensions, the multiscale GWR (MGWR) [6, 9] has many attractive features. MGWR specifics a unique bandwidth for each coefficient to improve the goodness of fit and prediction accuracy [9]. The hierarchical linear model [10], is also an important method for finding spatial heterogeneity in data of hierarchical structure. When samples are grouped by their locations, HLM calibrates some effects for samples in each group (called "random effects") to fit for spatially varying relationships, whereas other effects are treated as "fixed effects" that are constant for all groups [8].

In recent years, spatially hierarchical data have become increasingly popular in real world analysis since samples can be naturally nested in different spatial scales. For example, in the Biobank database [1] which consists of health information from 0.5 million UK participants, their addresses are nested into 1km-by-1km grid cells to protect their privacy. With the development of spatial big data and improved access to administrative data through secure data settings, it is increasingly common to find data sets where the attributes of the sample are available at a different geographic scale to their geographical identifiers. In spatial data of hierarchical structures, effects of variables may work in different ways. For example, group-level variables – that keep constant within groups – may have global or local effects, and sample-level variables are the same. No matter which variables, the basic GWR model always treat their effects as local ones, and estimate them by data borrowing from geographical neighbors. When dealing with group-level variables, the repeated values increase the risk of singular matrix. MGWR works similarly, only that it assigns variable-specified bandwidth settings and variables of global effects will be assigned a huge bandwidth up to infinity to estimate global effects. Fixed effects and random effects in HLM can be used to discover global and local effects, respectively. Fixed effects can be estimated for both group-level variables and sample-level variables. However, random effects only work for sample-level variables, which vary among individuals as opposed to the group-level ones. Because values of group-level variables are determined by their locations. Thus, there is no sufficient variation to calibrate random effects for them within each group. We need a special method to properly estimate effects of the variables with spatial heterogeneity.

In this article, we propose a hierarchical and geographically weighted regression (HGWR) model and its estimator based on backfitting and maximum likelihood (BFML) algorithms to solve the above-mentioned issues. This model calibrates two types of fixed effects – local fixed effects and global fixed effects – and random effects. We conducted a simulation experiment to ascertain whether HGWR could successfully distinguish local effects from other effects. We also compared its performance with GWR, MGWR, and HLM.

2 Model

The HGWR model is designed for data with a spatially hierarchical structure. In a data set with n samples divided in m groups according to their locations, the variance of dependent variable $\boldsymbol{y}$ can be explained with the following three parts: local-fixed effects $\boldsymbol{\gamma}$ for variables

$\boldsymbol{G}$ that vary with location; global-fixed effects $\boldsymbol{\beta}$ for variables X that are constant across the whole area; and random effects $\boldsymbol{\mu}$ for variables $\boldsymbol{Z}$ that vary from group to group. The model for sample j in group i can be expressed as Equation 1,

$$y_{ij} = \boldsymbol{G}_i\gamma_i + \boldsymbol{X}_i\beta + \boldsymbol{Z}_{ij}\boldsymbol{\mu}_i + \epsilon_{ij} \tag{1}$$

where $\boldsymbol{\gamma}_i$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\mu}_i$ are coefficients of local fixed effects, global fixed effects, and random effects respectively; $\boldsymbol{\epsilon}_{ij}$ is the remaining random error. Then this model can be written in a matrix from as Equation 2,

$$\boldsymbol{y} = \mathrm{diag}\left\{\boldsymbol{G\gamma}\right\} + \boldsymbol{X\beta} + \boldsymbol{Z\mu} + \epsilon \tag{2}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_m)$,

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_1 \\ \boldsymbol{G}_2 \\ \vdots \\ \boldsymbol{G}_m \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & & & \\ & \boldsymbol{Z}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{Z}_m \end{pmatrix} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_m \end{pmatrix},$$

$\boldsymbol{\mu} \sim \boldsymbol{N}(0, \sigma^2\boldsymbol{D})$, and $\boldsymbol{\epsilon} \sim \boldsymbol{N}(0, \sigma^2\boldsymbol{I})$, and $\boldsymbol{G\gamma}$ here is regarded as a product of block matrices, such that

$$\boldsymbol{G\gamma} = \begin{pmatrix} \boldsymbol{G}_1\gamma_1 & \boldsymbol{G}_1\gamma_2 & \cdots & \boldsymbol{G}_1\gamma_m \\ \boldsymbol{G}_2\gamma_1 & \boldsymbol{G}_2\gamma_2 & \cdots & \boldsymbol{G}_2\gamma_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_m\gamma_1 & \boldsymbol{G}_m\gamma_2 & \cdots & \boldsymbol{G}_m\gamma_m \end{pmatrix}, \mathrm{diag}\left\{\boldsymbol{G\gamma}\right\} = \begin{pmatrix} \boldsymbol{G}_1\gamma_1 \\ \boldsymbol{G}_2\gamma_2 \\ \vdots \\ \boldsymbol{G}_m\gamma_m \end{pmatrix}.$$

In this model, coefficients $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_m$ are estimated group-by-group as for other GWR models using weighted least squared estimation [3] with a uniform bandwidth.

A back-fitting procedure, shown in Figure 1, can be applied to estimate parameters in this model following a similar methodical approach to [4]. In this workflow, when calibrating local fixed effects $\hat{\boldsymbol{\gamma}}^{(t)}$ in each iteration, the algorithm can optimize the bandwidth value via golden-selection [7] according to the CV criterion. This algorithm is very efficient and effective in minimizing univariate functions.

3 Simulation Experiments

To evaluate the performance of HGWR and compare this model with HLM, GWR and MGWR, some simulation experiments [2] are designed. In particular, the performance was measured regarding the ability to properly distinguish local fixed effects from global fixed effects under the circumstance that random effects exist.

A spatial data set of 21,434 random samples was generated that were unevenly spread across 625 locations. The data generating process was inspired by [6]. For each data point, four independent variables $(\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{x}_1, \boldsymbol{z}_1)$ were generated according to the standard multivariate normal distribution. To simulate group-level spatial-related variables, the mean of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ at each location were substituted for the original values. Samples located together share coefficient values. Values of the generated coefficients are shown in the first row of Figure 2. Results of the for models are shown in other rows.

[2] Please turn to `https://hpdell.github.io/GIScience-Materials/posts/HGWR/` for more details.

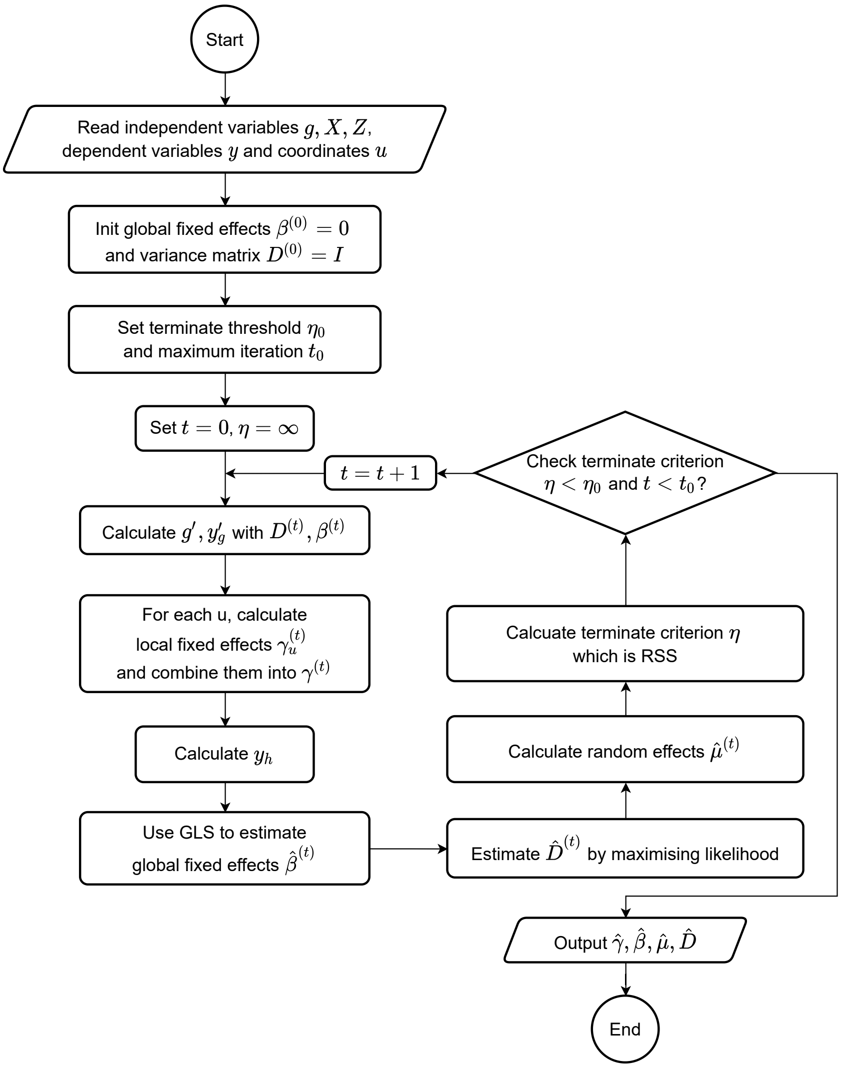

Figure 1 Diagram of the BFML estimator for HGWR, where RSS $= (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}})$ and $\hat{\boldsymbol{y}} = \boldsymbol{G}\hat{\gamma} + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\mu}}$.

In the results of GWR, spatial heterogeneity is revealed in estimates for all variables. Although $\hat{\boldsymbol{\beta}}_1$ should be constant across the study area, GWR still generate spatially varying estimates for it. This is a kind of over-fitting from the spatial perspective. However, for estimates of $\boldsymbol{\mu}_1$, they are smoothed compared with actual values, even though the bandwidth selected is small enough. Because the bandwidth is small, estimates for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are too local. Consequently, there are quite a few outliers disrupting the spatial trend.

MGWR partly gets over issues of GWR by adopting parameter-specified bandwidths, instead of a uniform bandwidth. It performs better when estimating $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. For global fixed effects, MGWR still generates spatially varying estimates, but they vary more slightly than estimates from GWR. For random effects, the results are slightly smoothed as well. Besides, MGWR it requires a lot of computing time and memory.

In the results of HLM, there is only one estimate for $\boldsymbol{\beta}_1$ across the whole area as well as estimates for $\boldsymbol{\mu}_1$, the problem lies in estimates for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. As they are fixed effects in HLM, their estimates are also constant for all samples. However, spatial heterogeneity is expected in them.

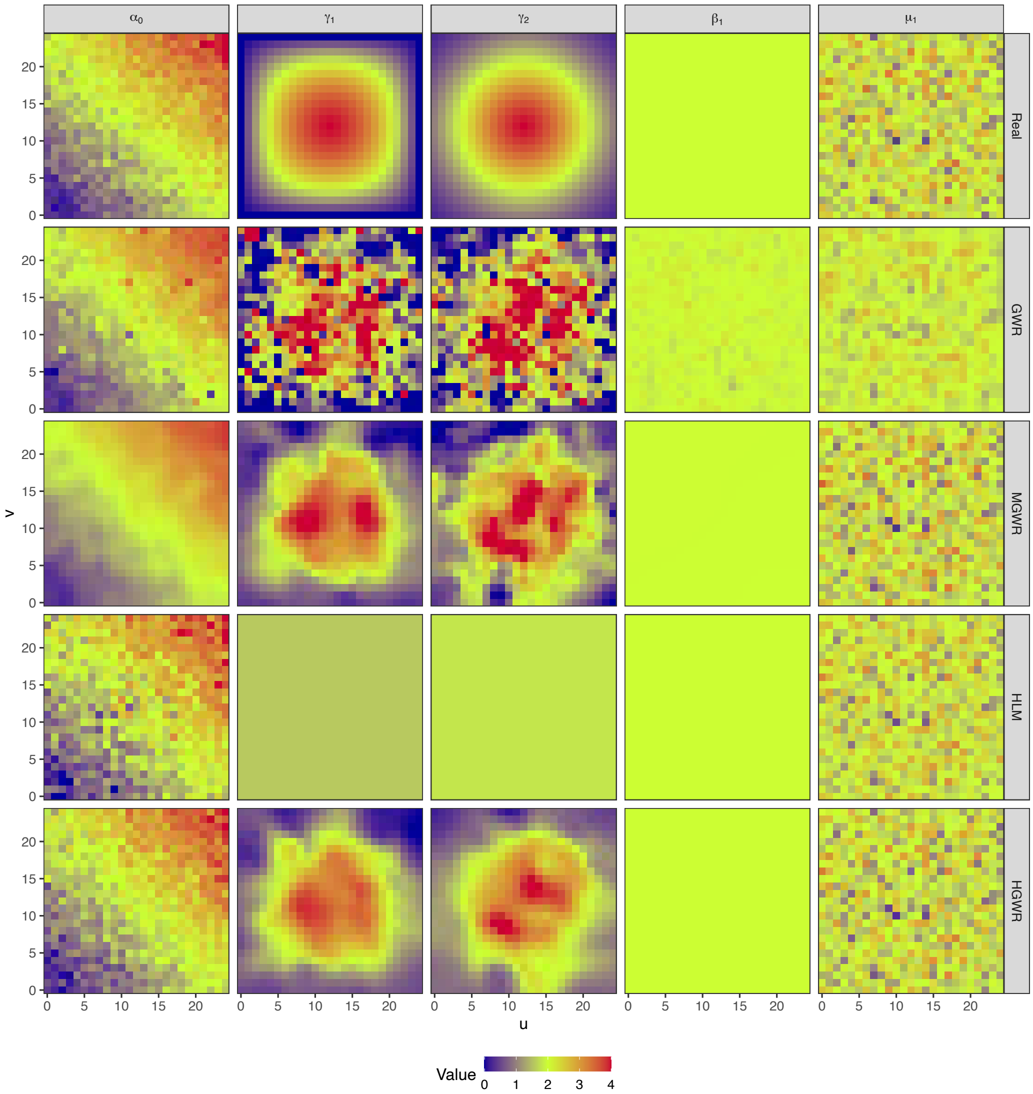

Figure 2 Real values and estimated values.

HGWR is the final solution. For global fixed effects, it generates globally constant estimates for all samples. For random effects, it does not smooth the estimates because they are not obtained by borrowing points. And for local fixed effects, we can discover spatial heterogeneity from their estimates. And it does not repeat computation for samples at each location. Computationally, it is more efficient because it does not repeat geographically weighted fitting at every sample within a higher-level group where models are the same. On the dataset used in the experiment, calibrating the HGWR model only took 6.06 seconds, which reduced the calculation time by 4 minutes compared to GWR (3.55 mins); and reduced it by nearly 4.4 hours compared to MGWR (4.41 hours) paralleled by 48 threads. These findings have been double-checked via repeating the experiment 100 times.

4 Conclusion

In this article we proposed a BFML estimator for a HGWR model. Compared with HLM, this method divides fixed effects into global and local effects. For local fixed effects, this model applies a spatial heterogeneity assumption and estimates the effects using the GWR method. For global fixed effects and random effects, this model adopts a similar method as in HLM, i.e., maximum likelihood. To facilitate cooperation between the two methods, a back-fitting procedure was developed. It is demonstrated that HGWR can properly estimate local fixed effects, global fixed effects, and random effects simultaneously. HGWR can successfully distinguish local fixed effects from other effect types. For local fixed effects, spatial heterogeneity is considered as with GWR; moreover, global fixed effects and random effects are estimated as accurately as when using HLM. Thus, HGWR can be regarded as a successful combination of GWR and HLM. In this stage, there are some limitations remaining to be solved, such as convergence conditions and statistical inferences. Nevertheless, with the popularity of spatiotemporal big data, situations wherein the specific parameters for which HGWR was optimized are becoming more prevalent, suggesting that HGWR holds considerable promise as a useful tool for analyzing such data sets.

References

1 Naomi E. Allen, Cathie Sudlow, Tim Peakman, and Rory Collins. Uk biobank data: Come and get it. *Science Translational Medicine*, 6(224):1–4, February 2014. `doi:10.1126/scitranslmed.3008601`.

2 Luc Anselin. *Spatial Econometrics: Methods and Models*. Number 4 in Studies in operational regional science. Kluwer Academic Publishers, Dordrecht, Boston, 1988.

3 Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, February 1996. `doi:10.1111/j.1538-4632.1996.tb00936.x`.

4 Feng Chen, Yee Leung, Chang-Lin Mei, and Tung Fung. Backfitting estimation for geographically weighted regression models with spatial autocorrelation in the response. *Geographical Analysis*, 54(2):357–381, April 2022. `doi:10.1111/gean.12289`.

5 A. Stewart Fotheringham and Chris Brunsdon. Local forms of spatial analysis. *Geographical Analysis*, 31(4):340–358, September 2010. `doi:10.1111/j.1538-4632.1999.tb00989.x`.

6 A. Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265, November 2017. `doi:10.1080/24694452.2017.1352480`.

7 Dorothy Margaret Greig. *Optimisation*. Longman Publishing Group, 1980.

8 A. M. Latimer, S. Banerjee, H. Sang Jr, E. S. Mosher, and J. A. Silander Jr. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern united states. *Ecology Letters*, 12(2):144–154, February 2009. `doi:10.1111/j.1461-0248.2008.01270.x`.

9 Binbin Lu, Chris Brunsdon, Martin Charlton, and Paul Harris. Geographically weighted regression with parameter-specific distance metrics. *International Journal of Geographical Information Science*, 31(5):982–998, May 2017. `doi:10.1080/13658816.2016.1263731`.

10 Stephen W. Raudenbush. Hierarchical linear models and experimental design. In *Applied Analysis of Variance in Behavioral Science*, pages 459–496. L. K. Edwards, 1993.