AGILE: GIScience Series, 5, 55, 2024. https://doi.org/10.5194/agile-giss-5-55-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Multilevel Geographical Process Models (MGPMs): A Novel Framework for Modeling Individual- and Multi-Level Spatial Process Heterogeneity

Sui Zhang [1] and Ziqi Li [2]

[1]Department of Planning, Property and Environmental Management, University of Manchester, Manchester, United Kingdom
[2]Department of Geography, Florida State University, Tallahassee, United States

Correspondence: Sui Zhang (sui.zhang@manchester.ac.uk)

Abstract. Scale, context, and heterogeneity have been central issues in geography. From a quantitative standpoint, accurately identifying the scale and context at which geographical processes operate and capturing their spatial heterogeneity have been challenging tasks. Despite various prominent developments in spatial modeling literature, there is a lack of models for separating individual- and group-level spatial processes that may also exhibit spatial heterogeneity. Understanding this difference can better inform us about how individuals are separated from or influenced by higher-level contexts. In this regard, we propose Multilevel Geographical Process Models (MGPMs) to simultaneously incorporate both individual- and multi-level spatial process heterogeneity. We demonstrate the performance of the model using Monte Carlo simulations and Compare it against Multiscale Geographically Weighted Regression and Multilevel models.

Keywords. multilevel models, multiscale geographically weighted regression, spatial process heterogeneity

1 Introduction

Over the past decades, understanding the role of geographical context, the scale at which geographical processes operate, and their heterogeneity has been of major interest to quantitative geographers. Statistical frameworks, including multilevel models (MLMs) and multiscale geographically weighted regression (MGWR) have seen common use in modeling contextual multiscale effects. MLMs are appropriate for hierarchical data, and heterogeneity among aggregated levels (which are defined a priori) can be modeled by including random effects (Kreft and de Leeuw, 1998). MGWR follows a local regression framework that allows parameters exhibit different levels of heterogeneity at the individual level (Fotheringham et al., 2017). Compared to MLMs, MGWR alleviates the possibility of modifiable areal unit problem (MAUP) by avoiding defining spatial regimes where processes are homogeneous within.

A recent study compared the two frameworks using both synthetic and empirical datasets (Fotheringham and Li, 2023). The results suggested that while MGWR and MLMs produced similar estimated spatial patterns of geographical contextual effects, MGWR indicated local varying processes with higher spatial resolution that cannot be fully explained by coarse upper hierarchical levels. However, does the detailed local effect arise from noise and model assumptions, or are they indeed interesting processes that should be identified but are missed by traditional MLMs? The question cannot be fully answered due to the lack of models that are able to accurately separate local- and group-level geographical processes. There are some efforts worth noting along this direction. Harris et al. (2013) reconstructed the GWR weight matrix leveraging contextual covariates to incorporate upper-level divisions into the model calibration. Dong and Harris (2015) generalised the paradigm of hierarchical spatial autoregressive models (HSAR) so that spatial dependencies at both the individual and group levels can be taken into account. Chen and Truong (2012) and Hu et al. (2022) combined MLM with GWR, proposing a hierarchical and geographically weighted regression (HGWR), which allows the spatially varying processes at the group level to be identified. Data-driven methods for regionalising the group-level divisions in MLM, based on local intercepts derived from GWR, have also been developed (Feuillet et al., 2024). Despite these developments, there is a gap in the literature to develop a unified modeling framework to achieve the following three objectives simultaneously:

- To model the individual-level spatially varying processes and explicitly provide the spatial scales at which they operate.
- To explicitly allow group-level covariates and model their group-level effects.
- To separate spatially varying processes at different hierarchies thereby enabling better process identification.

To address this gap, we propose multilevel geographical process models (MGPMs) as a novel paradigm for modeling hierarchical geographical data. This framework inherits the specification and the flexible structure from the traditional MLM but also further develops the superiority of MGWR in modeling geographical process heterogeneity. In this short paper, as a starting point, our discussion mainly focuses on the MGPM with fixed-slope and random-intercept regarding its specification and estimation, though it can be extended to consider random-slope. Through Monte Carlo simulations on a synthetic dataset, we demonstrate its advantages over traditional models in terms of parameter estimate accuracy, bandwidth estimation, and overall model accuracy and sensitivity.

2 Methods

2.1 Multilevel Geographical Process Model

A traditional MLM with fixed-slope and random-intercept can be represented in matrix notation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are vectors of fixed effects and random effects. $\mathbf{X}$ and $\mathbf{Z}$ represent design matrices for the fixed effects and random effects respectively. $\boldsymbol{\epsilon}$ is the error term. Such specification assumes that the β for each covariate is constant over space. In MGWR, by allowing the parameters to vary spatially at different scales, this assumption can be relaxed (but only in the single-level case):

$$y_i = \sum_{j=0}^{m} \beta_{bwj}\left(u_i, v_i\right) x_{ij} + \varepsilon_i \tag{2}$$

where x_{ij} is the jth explanatory variable, $\beta_{bwj}\left(u_i, v_i\right)$ is the jth coefficient, and the bwj indicates the bandwidth of the jth covariate. Leveraging the above techniques, we extend MLM to MGPM for modeling individual-level spatial process heterogeneity, which can be formulated as:

$$\mathbf{y} = \mathbf{X}_{(\mathbf{k=0})}\boldsymbol{\beta}_{(\boldsymbol{k}=\mathbf{0})} + \mathbf{X}_{(\mathbf{k>0})}\boldsymbol{\beta}_{(\boldsymbol{k}>\mathbf{0})} + \mathbf{Z}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{3}$$

where the subscript (k) denotes that this vector or matrix contains elements of the k levels, with $k = 0$ representing the level 1 (individual-level). However, here, the $\boldsymbol{\beta}_{(\boldsymbol{k}=\mathbf{0})}$ is no longer a vector, but a $j \times i$ matrix containing the individual-specific parameter $\beta_{ij(0)}$. When $bw0 = bw1 = \dots = bwj = N$ (N represents the total observations), the individual-level spatial processes can be considered to be spatially constant, the MGPM thus reduces to the MLM approximately (still with slight variations in the $\boldsymbol{\beta}_{(\boldsymbol{k}=\mathbf{0}}$ due to the weighting scheme in the MGWR). On the other hand, when $k = 0$, the MGPM reduces to the MGWR (random effects are not applicable in this case).

A back-fitting algorithm as follows can be adopted for the model calibration (Figure 1). In this study, the GWR bandwidths were optimised by Corrected Akaike Information Criterion (AICc), and the score of change in the residual sum of squares (SOC-RSS) $\leq 10^{-5}$ was set as the termination criterion of the calibration routines, according to Fotheringham et al. (2017).

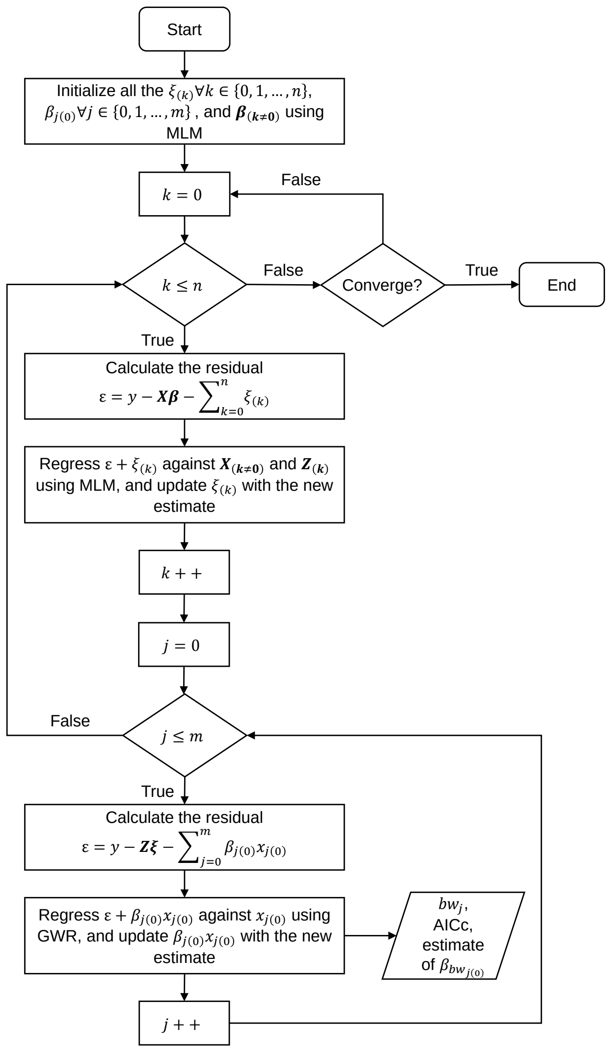

Figure 1. Back-fitting algorithm for an MGPM

2.2 Simulation Design

In order to compare the performance of MGPM with traditional models, a simulation experiment was designed. The simulated synthetic dataset was constructed based on the following data generating process (DGP):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \xi + \varepsilon \tag{4}$$

The simulated space was created over a regular 25 × 25 lattice. We then designed three different parameter surfaces for each of the individual-level covariates with zero, medium, and high spatial heterogeneity following Fotheringham et al. (2017) and a group-level random effect surface with medium spatial heterogeneity:

$$\beta_0 = 3 \tag{5}$$

$$\beta_1 = 3 + \frac{1}{12}(u - v) \tag{6}$$

$$\beta_2 = 1 + \frac{1}{324}\left[36 - \left(6 - \frac{u}{2}\right)^2\right]\left[36 - \left(6 - \frac{v}{2}\right)^2\right] \tag{7}$$

$$\xi = -2 + \frac{1}{12}(u + v) \tag{8}$$

where u and v is the horizontal and vertical coordinate respectively. The individual-level covariates x_1 and x_2 were draw randomly from a normal distribution $N(0,1)$ and the error term ε was generated from $N(0,0.5)$. We then repeated this process 100 times to examine the robustness of the results. The true parameter surfaces are visualized in Figure 2.

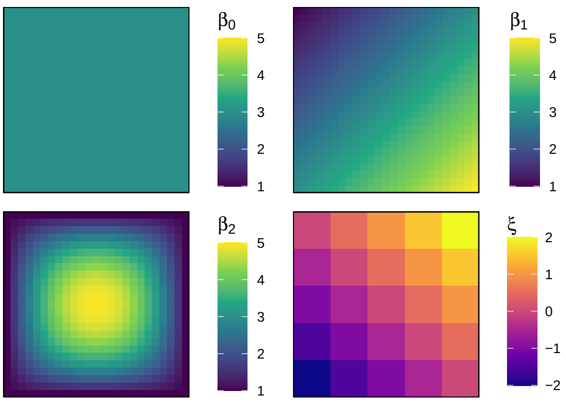

Figure 2. True parameter surfaces.

2.3 Data and Software Availability

All the data in this study was randomly generated by the R packages `base` and `stats`. R packages `lme4` and `GWmodel` were used to perform MLM and MGWR respectively. The R code used to perform MGPM will be developed as a package and released with the final outputs.

3 Results

3.1 Surface Recovery

Figure A1 shows the recovery performance from a single realisation of different models for the four parameter surfaces. Since the fixed-slope MLM follows the hypothesis that the individual-level processes are spatially constant, the spatial heterogeneity of both β_1 and β_2 was omitted, except for the zero heterogeneity intercept β_0. MGWR can estimate the spatially varying β_1 and β_2 accurately. However, it identified group-level random effects as an individual-level spatial varying process erroneously due to its incapacity in incorporating effects at upper hierarchies into the model calibration, thereby failing to recover β_0 vs ξ well. MGPM performed the best among all models in terms of surface recovery. It not only estimated the individual-level spatial varying processes β_1 and β_2 accurately, but also identified the group-level random effects and thus recovered the spatially constant β_1. It is remarkable that MGPM outperformed the traditional MLM in recovering ξ as the individual-level processes were estimated more accurately.

3.2 Bandwidth Comparison

Figure 3 shows the bandwidths derived from the MGWR and MGPM on the 100 simulated datasets (Bandwidth is not applicable for MLM). Both MGWR and MGPM can relatively correctly identify the spatial scales at which the β_1 and β_2 vary. However, the bandwidth estimate of β_0 from MGWR was extremely biased due to identifying the varying effects at the upper hierarchy as at the individual level, whereas MGPM estimated the scale at which the process operates accurately and robustly (with the median close to the maximum value of 625, implying a global process).

3.3 Parameter Estimates

Figure 4 illustrates the estimation accuracies of different models for the four parameter surfaces (ξ^* is not available for MGWR) on the 100 simulated datasets via root mean squared error (RMSE). The MLM had the worst estimation accuracies for the individual-level parameter surfaces β_1 and β_2 among all the models. Even surface with zero heterogeneity was not robustly estimated. MGWR was able to estimate both β_1 and β_2 relatively accurately,

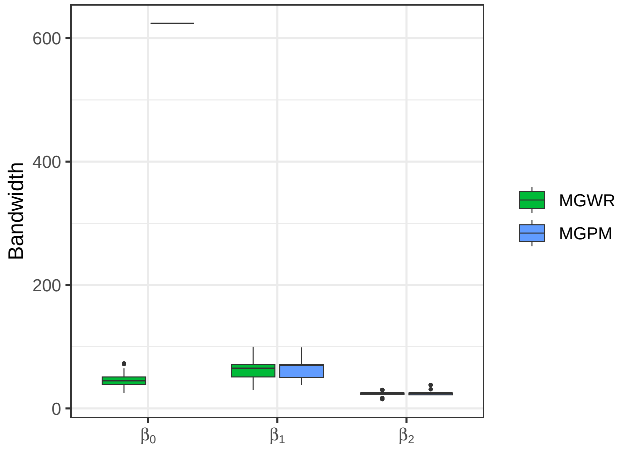

Figure 3. Optimal bandwidth from MGWR and MGPM.

but it performed the worst among the models in estimating β_0 due to the misidentification of the operation scale. The MGPM estimated every parameter surfaces the best, no matter at the individual level or at an upper hierarchy. Notably, MGPM even outperformed MGWR in the estimation of the individual-level spatially varying processes β_1 and β_2 as it considered multilevel spatial processes simultaneously. Meanwhile, its estimation of the group-level random effects was also found to be much more accurate and robust in comparison to the MLM.

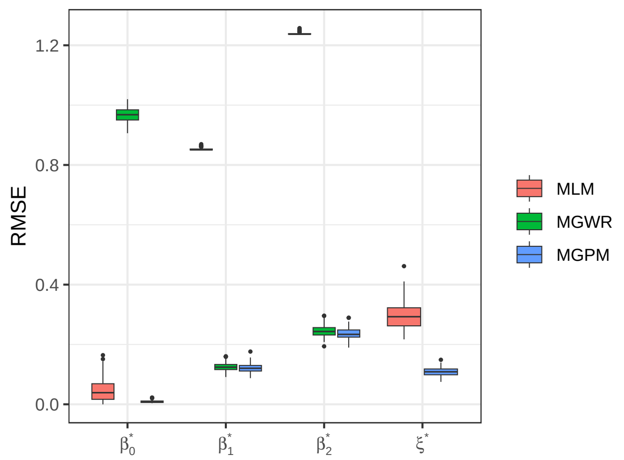

Figure 4. RMSE for each parameter surface from MLM, MGWR, and MGPM.

3.4 Goodness of Fit

Figure 5 shows the residual sum of squares (RSS) for different models on the 100 simulated datasets to compare their model fitting performance. MLM was the worst performing model due to its specification on fixed individual-level processes. The difference between MGPM and MGWR is not obvious whilst the MGPM is slightly superior. Given the results in Figure 4, this indicates that the MGPM can achieve better identification without harming the goodness of fit.

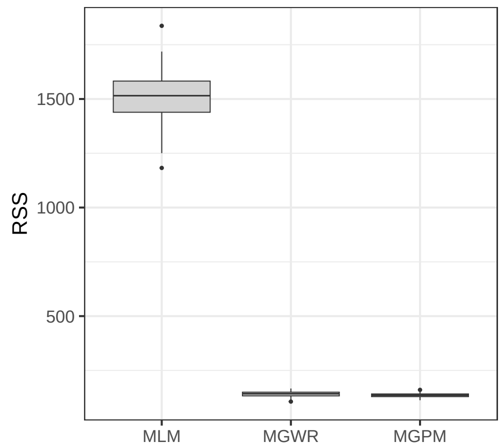

Figure 5. RSS for MLM, MGWR, and MGPM.

4 Discussion

Here we introduced the model specification and estimation of the MGPM with fixed-slope and random-intercept, and then demonstrated its modeling performance. The results of the Monte Carlo simulations suggested that the MGPM was the preferred model compared to both the traditional MLM and the MGWR in multiple aspects including parameter estimation and model fitting. Compared to the HGWR, the most recent advance, MGPMs have a more flexible framework: they are not restricted to a multilevel model specification with fixed-relation (random-slope and random-intercept), but also allow for individual-level geographical processes to be observed at the hierarchy at which they operate, rather than being aggregated to upper hierarchies (Hu et al., 2022).

Moreover, there are also promising applications for such an MGPM. For example, school performance is not only influenced by the characteristics of its catchment area geographically, but also by the administrative boundaries in which it is located (Fotheringham et al., 2001); Green space was indicated to have a spatial nonstationary impact on individual wellbeing, on which the effect of upper-level covariates such as the area level deprivation also worth to be further explored (Houlden et al., 2019); County-level voting preferences varied with a range of demographic characteristics, and they may also vary with the intrinsic context of the state they are in (Fotheringham et al., 2021).

However, there are some limitations of MGPMs at present: 1) the statistical inference framework that integrates multilevel and spatially varying models needs further investigation; and 2) the current implementation of MGPM does not support simultaneous random-slope and random-intercept specification, which will be developed in the next step of our work.

Competing interests. The authors declare that no competing interests are present.

References

Chen, D.-R. and Truong, K.: Using multilevel modeling and geographically weighted regression to identify spatial variations in the relationship between place-level disadvantages and obesity in Taiwan, Applied Geography, 32, 737–745, 2012.

Dong, G. and Harris, R.: Spatial autoregressive models for geographically hierarchical data structures, Geographical Analysis, 47, 173–191, 2015.

Feuillet, T., Cossart, E., Charreire, H., Banos, A., Pilkington, H., Chasles, V., Hercberg, S., Touvier, M., and Oppert, J. M.: Hybridizing Geographically Weighted Regression and Multilevel Models: A New Approach to Capture Contextual Effects in Geographical Analyses, Geographical Analysis, 2024.

Fotheringham, A. S. and Li, Z.: Measuring the Unmeasurable: Models of Geographical Context, Annals of the American Association of Geographers, 113, 2269–2286, 2023.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C.: Spatial variations in school performance: a local analysis using geographically weighted regression, Geographical and environmental Modelling, 5, 43–66, 2001.

Fotheringham, A. S., Yang, W., and Kang, W.: Multiscale geographically weighted regression (MGWR), Annals of the American Association of Geographers, 107, 1247–1265, 2017.

Fotheringham, A. S., Li, Z., and Wolf, L. J.: Scale, context, and heterogeneity: A spatial analytical perspective on the 2016 US presidential election, Annals of the American Association of Geographers, 111, 1602–1621, 2021.

Harris, R., Dong, G., and Zhang, W.: Using Contextualized Geographically Weighted Regression to Model the Spatial Heterogeneity of Land Prices in Beijing, China, Transactions in GIS, 17, 901–919, 2013.

Houlden, V., de Albuquerque, J. P., Weich, S., and Jarvis, S.: A spatial analysis of proximate greenspace and mental wellbeing in London, Applied Geography, 109, 102 036, 2019.

Hu, Y., Lu, B., Ge, Y., and Dong, G.: Uncovering spatial heterogeneity in real estate prices via combined hierarchical linear model and geographically weighted regression, Environment and Planning B: Urban Analytics and City Science, 49, 1715–1740, 2022.

Kreft, I. and de Leeuw, J.: Introducing Multilevel Modeling, Introducing Statistical Methods, SAGE Publications, Ltd, London, https://doi.org/10.4135/9781849209366, 1998.

Appendix A

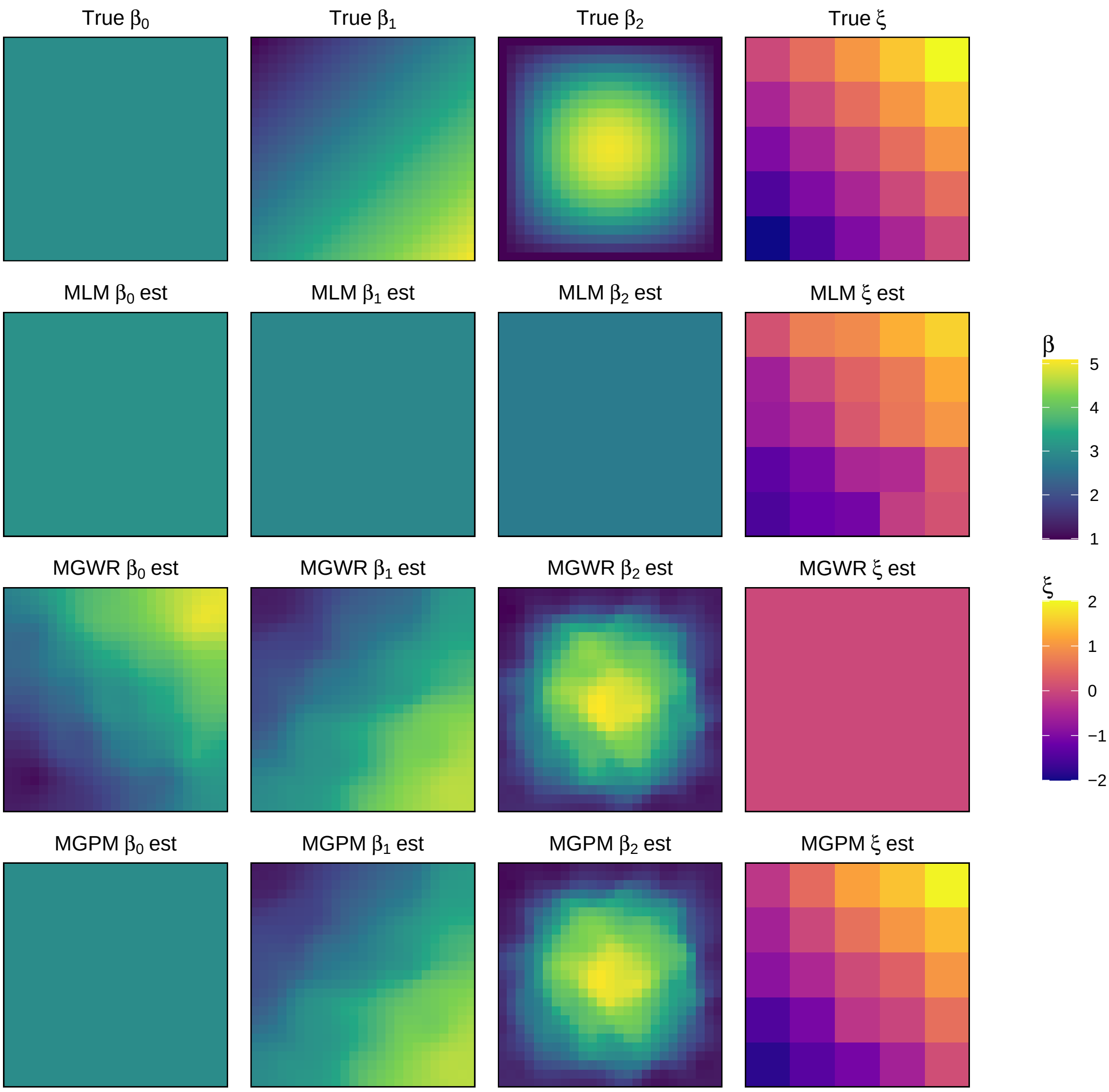

Figure A1. Recovered parameter estimate surfaces from a single realisation of MLM, MGWR, and MGPM.