



Introducing a General Framework for Locally Weighted Spatial Modelling Based on Density Regression

Yigong Hu¹  

School of Geographical Sciences, University of Bristol, UK

Binbin Lu  

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

Richard Harris  

School of Geographical Sciences, University of Bristol, UK

Richard Timmerman  

School of Geographical Sciences, University of Bristol, UK

Abstract

Traditional geographically weighted regression and its extensions are important methods in the analysis of spatial heterogeneity. However, they are based on distance metrics and kernel functions compressing differences in multidimensional coordinates into one-dimensional values, which rarely consider anisotropy and employ inconsistent definitions of distance in spatio-temporal data or spatial line data (for example). This article proposes a general framework for locally weighted spatial modelling to overcome the drawbacks of existing models using geographically weighted schemes. Underpinning it is a multi-dimensional weighting scheme based on density regression that can be applied to data in any space and is not limited to geographic distance.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Spatial heterogeneity, Multidimensional space, Density regression, Spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.40

Category Short Paper

Supplementary Material *Software (Source code)*: <https://github.com/GWmodel-Lab/GWmodel13>
archived at `swh:1:dir:24841fa8fac1919085decceb53131f35634b6b01`

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

1 Introduction

In recent years, analysis of spatial heterogeneity – for example, spatially varying regression relationships – has attracted increasing interest from researchers. Among the local-form spatial modelling methods, geographically weighted regression (GWR) [1] is popular. It fits a unique weighted least squared model at multiple locations across a study region by borrowing points from each location’s geographic neighbours. Extensions include geographically and temporally weighted regression (GTWR) [2], enhancing basic GWR’s ability to model more kinds of data. Basic GWR, on 2D spatial data sets, uses weights based on geographic distances between samples. Extended versions may adapt the weights to incorporate other

¹ Corresponding author.



kinds of “distance” but are still rooted back into one-dimensional distance metrics. This raises the problem of how to compress differences in multidimensional coordinates into a one-dimensional distance value.

Additionally, even when the metric is simple, differences in geographic scales of different dimensions may cause unexpected problems. This phenomenon is called “anisotropy”. For example, the range of vertical distances is generally different from that of horizontal distances. Consequently, when we incorporate distance in the 3D space to weight samples, relatively large changes in heights may present very limited effects on weights (without rescaling the vertical distances, at least). The problem is more evident when time is considered as this is, of course, measured in units of time, not of space. They are not directly compatible. These problems highlight the limitation of reducing multidimensional spaces into a single-dimensional weighting based on some notion of “closeness” or least distance.

In this paper, we introduce a general framework for locally weighted geographic and other spatial modelling based on density regression (DLSM: density-based local spatial models). This model essentially follows the workflow of density regression [6] under a conditional variable, but the conditional variable is restricted to the multivariate coordinates of samples in their space. Critically, this space can be geographic, spatio-temporal, or any other kind. It can have a dimension of any positive integer. Assuming these dimensions are independent, the DLGM framework calculates a weight for each according to their own bandwidth and kernel function. The product of these weights is used as the final weight to calibrate the least-squared model at each location. This modelling method can be easily adapted to any data of coordinates without trying to collapse the multiple dimensions into a single distance metric in the first instance. Simulation experiments demonstrate that this method is flexible, extensible and customisable. It can also reach higher goodness of fit than specially designed GWR-like models that attempt to accommodate spaces and coordinate systems that are not solely geographical.

2 Methodology

Geographically weighted regression can be expressed as Equation 1 for the sample i at location \mathbf{u}_i ,

$$y_i = \beta_{0i}(\mathbf{u}_i) + \beta_{1i}(\mathbf{u}_i)x_{1i} + \beta_{2i}(\mathbf{u}_i)x_{2i} + \cdots + \beta_{pi}(\mathbf{u}_i)x_{pi} + \epsilon_i \quad (1)$$

and the estimator for its coefficients $\beta_i = (\beta_{0i}, \beta_{2i}, \cdots, \beta_{pi})$ is shown in Equation 2,

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

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the vector of dependent variables, n is the number of samples, \mathbf{X} is the design matrix or independent matrix of all independent variables, $\epsilon_i \sim N(0, \sigma^2)$ is the random error and \mathbf{W}_i is the geographical weighting matrix for this sample. This weighting matrix is a $n \times n$ diagonal matrix. Each diagonal element is a distance-decay weight $w_{ij} = k(d_{ij}; b)$ (for $j = 1, 2, \cdots, n$) in which d_{ij} is the distance from sample i to j , k is a kernel function and b is the bandwidth. The basic GWR model uses straight-line distance, Minkowski distance, network distance, or travel time [4], which are all spatial. The GTWR model uses the spatial-temporal distance $d_{ij}^{ST} = d_{ij}^S \oplus d_{ij}^T$ by combining spatial distance and temporal distance together [2]. The bandwidth can be fixed (defined by distance), or adaptive (defined by the number of nearest neighbours).

For DLSM, the weight w_{ij} originates as a product of weights for every dimension in the current space, as shown in Equation 3,

$$w_{ij} = \prod_{h=1}^m w_{ijh} = \prod_{h=1}^m k_h(d_{ijh}; b_h) \quad (3)$$

where m is the number of dimensions in \mathbf{u}_i , k_h is the kernel function for dimension h , b_h is the corresponding bandwidth, $d_{ijh} = |u_{ih} - u_{jh}|$, and u_{ih}, u_{jh} is the coordinates in this dimension of sample i and j . Regardless of whether they are measured as longitude, latitude, height, time, social distance or any other measure of “closeness”, they are all feasible dimensions in this model. The estimator of this model can be that shown in Equation 2 or another locally weighed regression estimator.

The weighting method shown in Equation 3 operationalises multiple values of bandwidths – one for each dimension of the various coordinate spaces. The optimization of these bandwidths uses multidimensional minimisation of a criterion function. Theoretically, any kinds of multidimensional minimizer without derivatives are applicable here. We choose the Nelder-Mead algorithm [5]. The criterion function can be either the cross-validation (CV) value or goodness-of-fit, e.g., AIC function of given bandwidth $\mathbf{b} = (b_1, b_2, \dots, b_m)$, shown in Equation 4 and Equation 5 respectively,

$$\text{CV}(\mathbf{b}) = \sum_{i=1}^n [y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{-i}(\mathbf{b})]^2 \quad \text{or} \quad \text{CV}(\mathbf{b}) = \sum_{i=1}^n |y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{-i}(\mathbf{b})| \quad (4)$$

$$\text{AIC}(\mathbf{b}) = 2n \ln \hat{\sigma} + n \ln 2\pi + n \left[\frac{n + \text{tr}(\mathbf{S})}{n - 2 - \text{tr}(\mathbf{S})} \right] \quad (5)$$

where $\hat{\boldsymbol{\beta}}_{-i}(\mathbf{b})$ is the coefficient estimates for sample i without the sample itself, \mathbf{x}_i is the i -th row of matrix \mathbf{X} , \mathbf{S} is the “hat matrix” in which each row \mathbf{s}_i equals to $\mathbf{x}_i(\mathbf{X}\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}\mathbf{W}_i$.

3 Experiments

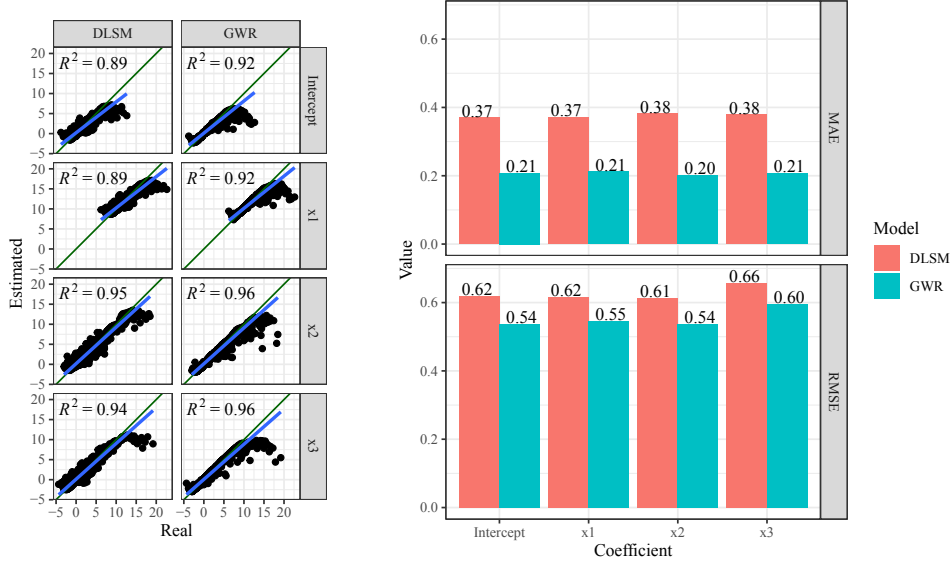
We carried out three experiments, generating simulation data sets to demonstrate how DLSM works². We also calibrated a corresponding GWR-family model in each experiment to provide a comparison. In each experiment, we use root mean squared error (RMSE) or mean absolute error (MAE) to evaluate the precise of estimates, which are defined in Equation 6,

$$\text{RMSE} = \sum_{i=1}^n (r_i - e_i)^2, \quad \text{MAE} = \sum_{i=1}^n |r_i - e_i| \quad (6)$$

where n is the number of estimates, e_i is the i -th estimate, r_i is the corresponding real value.

We first generated a 2D data set of Cartesian coordinates. Anisotropy was preserved in the coefficients. Bandwidths optimized by DLSM are 11.4% (570 neighbours) in the E-W direction and 0.7% (35 neighbours) in the N-S direction. Coefficient estimates and their RMSEs are shown in Figure 1. Whereas DLSM helps identify anisotropy, it is missing in estimates from a basic GWR model because the only bandwidth value optimized by GWR is 16 nearest neighbours (regardless of direction). It also has a stronger risk of overfitting as the bandwidth is too small. By contrast, DLSM can restrain overfitting in dimensions where spatial heterogeneity is weaker.

² Please turn to <https://hpdell.github.io/GIScience-Materials/posts/DLSM/> for more details.



(a) Comparison between estimates and (b) RMSE and MAE of coefficient estimates. real values.

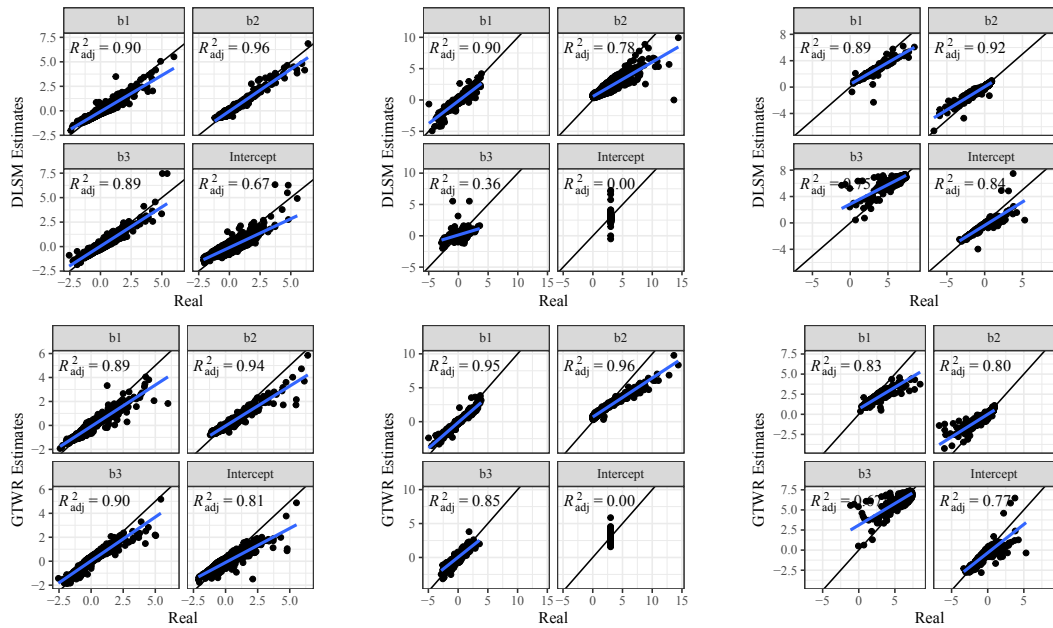
■ **Figure 1** Results of GWDR and basic GWR with two-dimensional spatial data.

Four 3D data sets of Cartesian coordinates representing space-time location (u_1, u_2, u_3) were generated to compare DLSTM and GTWR. In the former two data sets, coefficients were generated by $\exp(u_3)$. While in the latter two data sets, an autoregression model on u_3 was a part of all coefficients. The space-time distance metric use by GTWR was set to $d_{ij}^{ST} = \sqrt{\lambda(\Delta u_{1,ij}^2 + \Delta u_{2,ij}^2) + \mu(\Delta u_{3,ij}^2)}$. Parameters λ and μ in this space-time distance metric were optimized according to goodness of fit. Coefficient estimates and their RMSEs are shown in Figure 2. According to the results, DLSTM can reduce the mean of absolute estimation error by 10%-50%, especially when coefficients are temporally autocorrelated. The multiple bandwidths attach actual meaning to the parameters λ, μ ; they have a real-world correlate, unlike the root of sum of squared meters and seconds ($\sqrt{m^2 + s^2}$).

A 4D data set was also generated to simulate flow data. DLSTM was compared with GWR. For flow data, each flow can be represented by a set of 4D coordinates (u, v, α, l) in which u, v represents the spatial location of its starting point, α represents its direction, and l represents its length. The distance metric used by GWR was set to the similarity between flows $O_i(u_{O_i}, v_{O_i}) \rightarrow D_i(u_{D_i}, v_{D_i})$ and $O_j(u_{O_j}, v_{O_j}) \rightarrow D_j(u_{D_j}, v_{D_j})$ [3], as shown in

$$d_{ij} = \sqrt{\frac{[(u_{O_i} - u_{O_j})^2 + (v_{O_i} - v_{O_j})^2] + [(u_{D_i} - u_{D_j})^2 + (v_{D_i} - v_{D_j})^2]}{l_i l_j}} \quad (7)$$

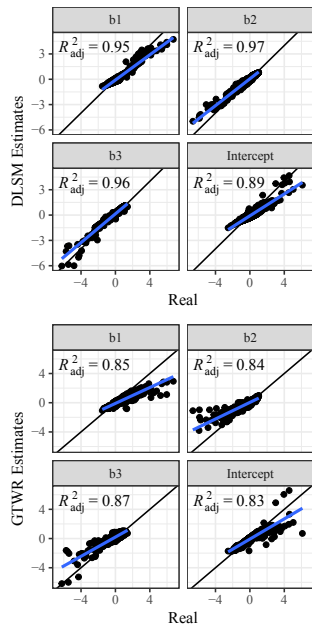
in which l_i is the length of flow $\overrightarrow{O_i D_i}$. Coefficient estimates and their RMSEs are shown in Figure 3. Results show that DLSTM works well for spatial line data even without defining distance metrics. It performs better than GWR according to the mean of estimation errors, but a few outliers exist in estimates. GWR selected a much smaller bandwidth (173 neighbours). Thus, the risk of overfitting reappears.



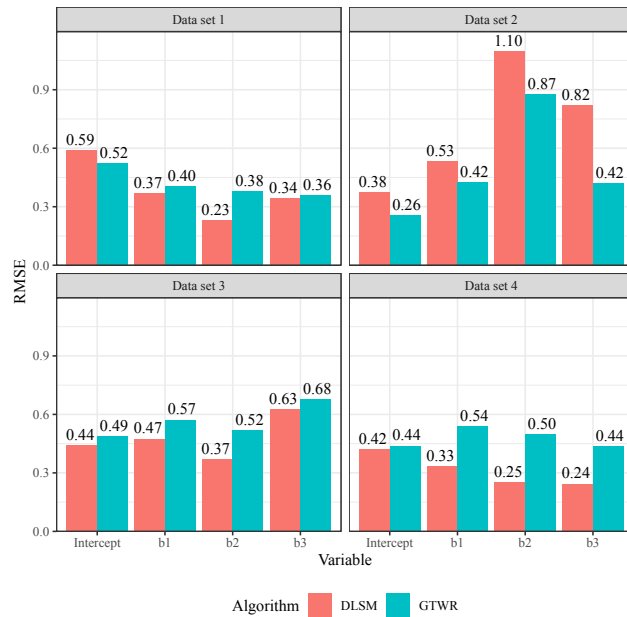
(a) Coefficient estimates and real values, the first data set.

(b) Coefficient estimates and real values, the second data set.

(c) Coefficient estimates and real values, the third data set.

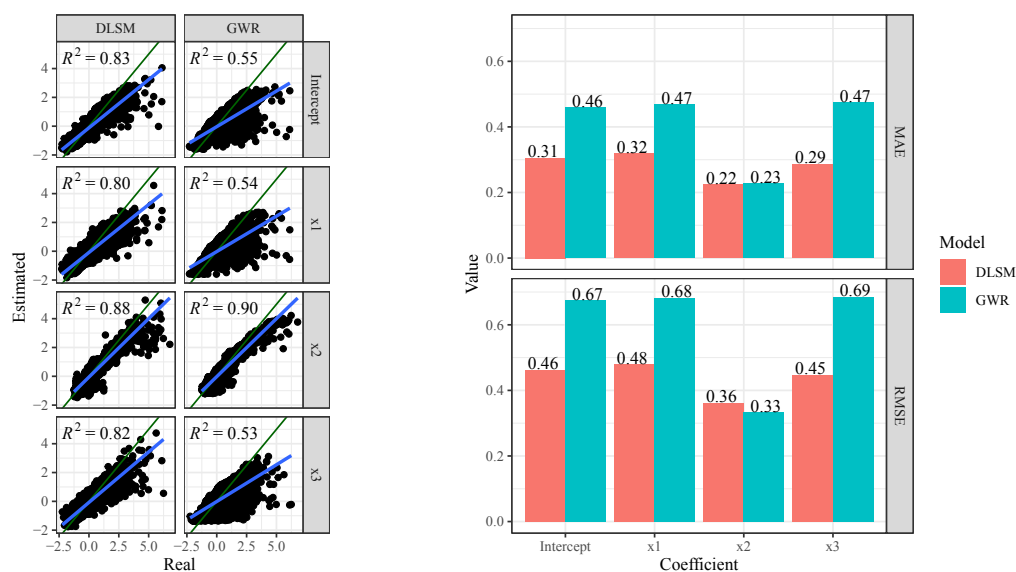


(d) Coefficient estimates and real values, the fourth data set.



(e) RMSE of estimates for each coefficient from DLSM and GTWR models on each data set.

Figure 2 Comparison between real value and estimations of coefficients given by GWDR and GTWR for ordinary spatial and temporal data.



(a) Comparison between estimates and real values. (b) RMSE and MAE of coefficient estimates.

■ **Figure 3** Results of GWDR and basic GWR with four-dimensional spatial data.

4 Conclusion

This paper introduces the DLSTM model as a framework for estimating local regression models, such as GWR and GTWR. It offers more flexibility because of its three alterable parts: a space where samples exist, a set of kernels selected for every dimension and a locally weighted regression method. Simulation shows that DLSTM can be applied to many kinds of spatial data without specially defined distance metrics, such as spatio-temporal data and spatial interaction data. It can also help tackle the effects of anisotropy because it has, in effect, a multidimensional bandwidth and decay function, measuring “closeness” in multiple dimensions simultaneously. In the future, researchers no longer need to design distance metrics to bring together, in a rather ad hoc way, different types of space and coordinate systems into the distance decay function. Assigning a weighting scheme to each of the dimensions and then pooling across them is suggested as a better alternative.

References

- 1 Chris Brunson, 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.
- 2 Bo Huang, Bo Wu, and Michael Barry. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401, March 2010. doi:10.1080/13658810802672469.
- 3 Maryam Kordi and A. Stewart Fotheringham. Spatially Weighted Interaction Models (SWIM). *Annals of the American Association of Geographers*, 106(5):990–1012, September 2016. doi:10.1080/24694452.2016.1191990.
- 4 Binbin Lu, Martin Charlton, Paul Harris, and A. Stewart Fotheringham. Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house

- price data. *International Journal of Geographical Information Science*, 28(4):660–681, April 2014. doi:10.1080/13658816.2013.865739.
- 5 J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965. doi:10.1093/comjnl/7.4.308.
- 6 Geoffrey S Watson. Smooth Regression Analysis. *The Indian Journal of Statistics, Series A*, 26(4):359–372, December 1964.