Geographically Weighted Regression Using a Non-Euclidean Distance Metric with Simulation Data

Binbin Lu Martin Charlton Paul Harris
National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co.kildare, Ireland
Contact: tel.: +353-1-7086731; fax:+353-1-7086456; E-mail: binbin.lu.2009@nuim.ie

Abstract—**In this study, we investigate the performance of a non-Euclidean distance metric in calibrating a Geographically Weighted Regression (GWR) model with a simulated data set. Random predictor variable and spatially varying coefficients are generated on a square grid of size 20*20. We respectively apply Manhattan and Euclidean distance metrics for the GWR calibrations. The preliminary findings show that Manhattan distance performs significantly better than the traditional choice for GWR - Euclidean distance. In particular, it out-performs in the accuracy of coefficient estimates.**

Keywords - Geographically Weighted Regression, non-Euclidean distance, Manhattan distance, simulation data

I. INTRODUCTION

In the early development of spatial analytical techniques in quantitative geography, invariably the techniques were applied at a 'global' level, where relationships were assumed to be constant across the study region [1]. However, the existence of spatial non-stationarity that appears in reality as uncontrolled spatial variability, challenges these global methods. This concept can be considered to be a particular form of the second law of Geography in the principle of spatial heterogeneity or non-stationarity [2]. In recent years, there has been an increasing interest in local multivariate methods for spatial data analysis that produce local results instead of 'one-size-fits-all' results from traditional global methods [1, 3]. This evolution is well reflected in the development of a particular class of spatial regression methods, where a number of local regression techniques have been introduced which estimate relationships that vary across space [4]. In Tobler's first law (TFL) [5, p236], he indicated that "Everything is related to everything else, but near things are more related than distant things". Inherited from TFL, Geographically Weighted Regression (GWR) has been developed as an important local technique to investigate spatial non-stationarity in data relationships [6].

The TFL statement requires a clear understanding of the meaning of "near". Accordingly, the concept of distance gives a quantitative or qualitative description of the nearness or similarity between any pair of objects or entities. In discussions on how it impacts on "related" objects, distance is generally recognised as an enervating factor that attenuates spatial interaction [7]. For GWR, the concern is to model a hypothetical 'bump of influence' that surrounds each regression point, where nearer observations are given more influence in estimating a set of local regression coefficients than observations farther away. In particular, a spatial weighting function is incorporated into its calibration to represent the influence of "near" observations in each "related" location-specific regression estimation. In practice, Euclidean distance (ED) is used to calculate the weighting matrix for GWR, although great circle measurements for un-projected geographical coordinates are also used [8]. However, the scope of possible distance metrics in spatial analysis is far larger than simple Euclidean ones. Due to our incomplete knowledge of geographical space, it is a complex system rather than an intuitive "table-top space" [9]. Its complexity determines that there can be no globally imposed distance metric for spatial analysis.

The many complex and diverse situations that GWR can be applied requires the option of specifying a non-ED metric in addition to the Euclidean one. In some situations, an untested usage of the ED metric may lead to inaccurate coefficient estimates, and in turn, a spatial pattern in estimates that is wrongly interpreted, due to artifacts from the straight-line measure between regression and data points. Intuitively, the more appropriate the distance metric is, the better GWR should perform. Lu et al [10] has applied network distance on modelling London House Price Data and the results have shown significant improvements on GWR. However, such empirical work produced a few difficulties in exactly exploring how a non-Euclidean metric works for GWR in terms of accuracy and effects control. The performances of implementing non-Euclidean distance metrics in GWR cannot be fully addressed due to the uncertainties caused by those defects. As such, simulated data is used to investigate performances of non-Euclidean distance metrics in calibrating GWR models in this paper.

II. Methodology

A. Using non-Euclidean distance metrics in GWR

A general form of a basic GWR model can be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \qquad (1)$$

where y_i is the dependent variable at location *i*; x_{ik} is the value of the *k*th independent variable at location *i*; *m* is the number of independent variables; $\beta_0(u_i, v_i)$ is the intercept parameter at location *i*; $\beta_k(u_i, v_i)$ is the local regression coefficient for the *k*th independent variable at location *i*; (u_i, v_i) are the coordinates of location *i*; and ε_i is the random error at location *i*.

The matrix expression for the estimation of the above model can be expressed as

$$\hat{\beta}(u_i, v_i) = \left(X^T W(u_i, v_i) X\right)^{-1} X^T W(u_i, v_i) y \qquad (2)$$

where *X* is the matrix of the independent variables with a column of *1*s for the intercept; y is the dependent variable vector; $\hat{\beta}(u_i, v_i) = \left(\beta_0(u_i, v_i), \cdots, \beta_n(u_i, v_i)\right)^T$ is the vector of *n+1* local regression coefficients; and $W(u_i, v_i)$ is the diagonal matrix denoting the geographical weighting of each observed data for regression point *i*. Here, the weighting scheme $W(u_i, v_i)$ is calculated with a kernel function based on the proximities between regression point *i* and the *n* data points around it. In practice, Euclidean distance (ED) is generally employed with planar coordinates.

The Akaike Information Criterion (AIC) [11], derived from the Kullback-Liebler information distance (KLID) [12] is used as the model diagnostic, it measures both goodness-of-fit and degrees of freedom. In practice, a corrected version of the AIC is applied for both model fit and bandwidth selection, and its expression is shown as [13]:

$$\text{AIC}_c(b) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n + \text{tr(S)}}{n - 2 - \text{tr(S)}}\right\} \qquad (3)$$

where *b* is the kernel bandwidth. The bandwidth is the key controlling parameter for GWR calibration and can be specified either by a fixed distance or by a fixed number of nearest neighbours, which respectively refers to two terms: fixed spatial kernel and adaptive spatial kernel.

Observe that the distance metric is an essential but separate component of the GWR technique. There is no special statement that the distance metric d_{ij} has to be Euclidean. Thus, the ED metric can be directly replaced by an appropriate non-ED measure in the basic GWR model (with a proviso the matrix algebra remains valid). The theoretical framework of GWR and related GW models [e.g. GWPCA - see 14] can still be followed with a generalized distance metric (Euclidean or non-Euclidean). In the next section, simple experiments are performed to exemplify this usage of a non-ED in GWR.

B. Simulation design

Currently, only a simple simulation has been conducted to provide some preliminary results. In this simulation, a data set of size 20*20 is generated on a square grid. For these data points, a predictor variable x_1 is generated as a random numeric vector ranging from 1 to 100, as shown in figure 1.

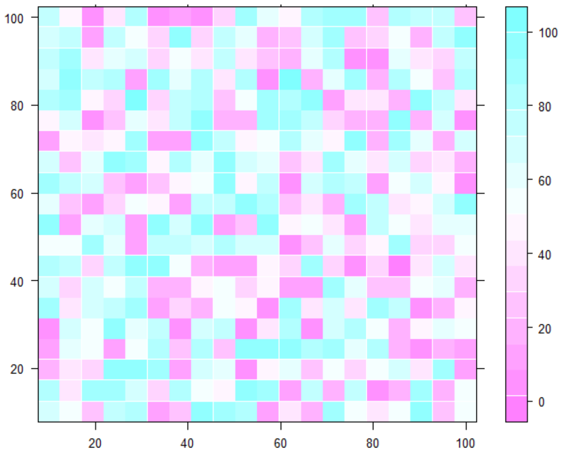


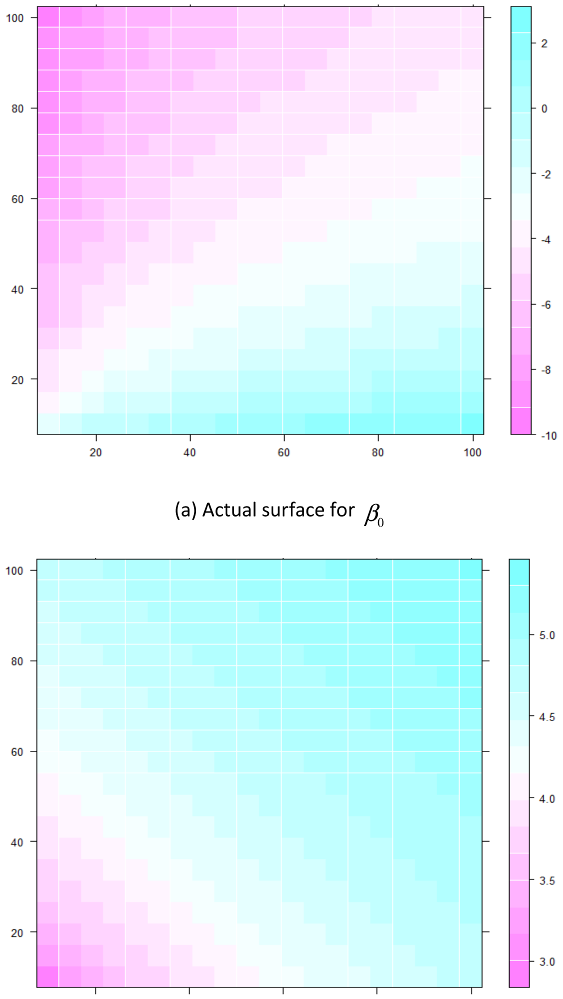
Figure 1 Generated predictor variable *x₁*

$$\beta_0(u_i, v_i) = 2\log(u_i) - 3\log(v_i) \qquad (4)$$

$$\beta_1(u_i, v_i) = \log(u_i + v_i) \qquad (5)$$

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) x_{1i} \qquad (6)$$

Two non-stationary regression coefficient surfaces, i.e. β_0 and β_1 are also generated by following equations (4) and (5). The actual surfaces are shown in the figure 2. Accordingly, the dependent variable is generated by following the basic linear model (formula (6)), as shown in figure 3.



(a) Actual surface for β_0

(b) Actual surface for β_1

Figure 2 Non-stationary regression coefficient surfaces for β_0 and β_1

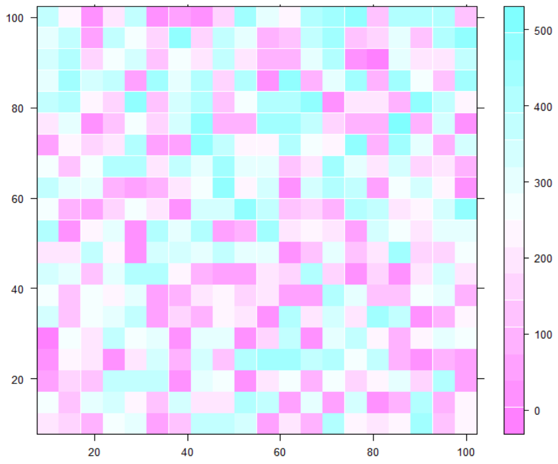


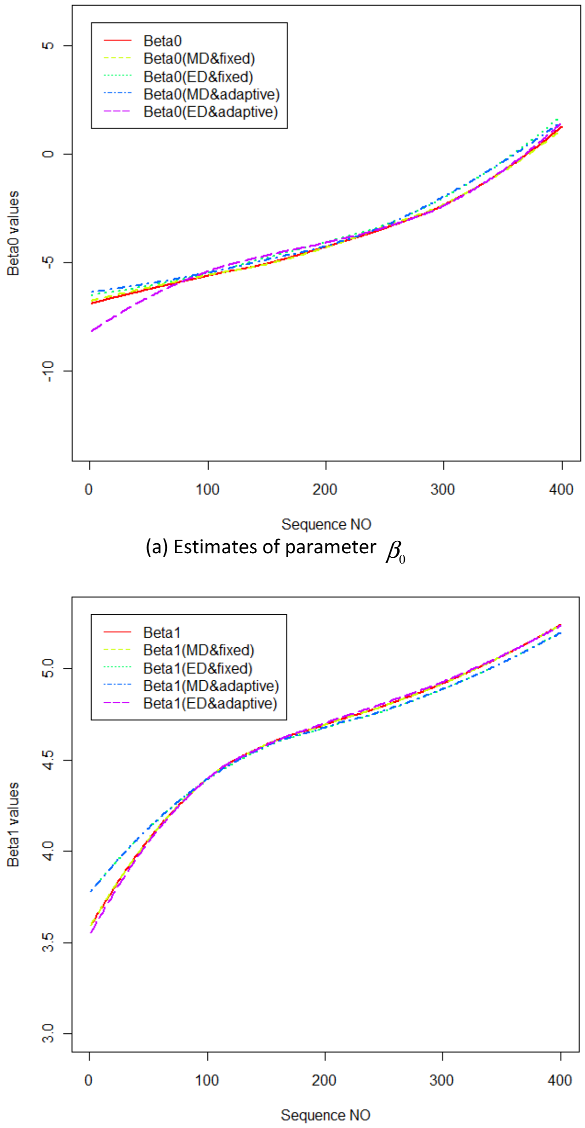
Figure 3 Generated dependent variable *y*

III. . Results

The data set is simulated on a regular grid, and then the Manhattan distance (MD) is tried for calibrating the above model. In these tests, the AICc values are used for both model diagnostic and bandwidth selection. Moreover, both fixed and adaptive spatial kernels are tested respectively with ED and MD.

TABLE I. Diagnostic information of the OLS and GWR calibrations using ED and MD with fixed and adaptive kernels

	OLS	**MD& fixed**	**MD& adaptive**	**ED& fixed**	**ED& adaptive**
AICc	3715.6	1616.1	2425.3	1641.0	2446.9



(a) Estimates of parameter β_0

(b) Estimates of parameter β_1

Figure 4 Coefficient (β_0 and β_1) and their estimates in the four calibrations

The AICc values of the OLS and four GWR calibrations are shown in table I. Overall, the GWR calibrations perform much better than OLS; the calibrations with fixed spatial kernels show significant improvements over those with adaptive kernels. Particularly, the calibration using MD with a fixed spatial kernel makes the best performance according to the smallest AICc value. This is also clearly reflected by the comparisons between the actual coefficients and corresponding estimates from these calibrations. Seen from figure 4, the coefficient estimates (for β_0 and β_1) from the calibration using MD with a fixed spatial kernel demonstrate the best approximation to their actual values. In summary, MD has shown a better fitting performance over ED in this simulated study.

IV. Conclusion

In this paper, a simple simulation study is conducted to investigate performances of ED and MD in calibrating a GWR model. The preliminary results have displayed a promising outlook of applying non-ED metrics in calibrating GWR models. Currently, more work is being undertaken to test our regression models and more fully understand the performance of a non-ED metric in GWR. This includes incorporating different weighting schemes and specifying more realistic simulations (e.g. those that include barriers where ED and non-EDs are clearly different between two data locations).

ACKNOWLEDGMENTS

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan.

REFERENCES

[1] A. S. Fotheringham and C. Brunsdon, "Local Forms of Spatial Analysis," Geographical Analysis, vol. 31, pp. 340-358, 1999.

[2] M. F. Goodchild, "The Validity and Usefulness of Laws in Geographic Information Science and Geography," Annals of the Association of American Geographers, vol. 94, pp. 300-303, 2004.

[3] A. Páez, "Local Analysis of Spatial Relationships: A Comparison of GWR and the Expansion Method," in Computational Science and Its Applications – ICCSA 2005. vol. 3482, O. Gervasi, et al., Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 631-637.

[4] D. Gamerman, et al., "Space-varying regression models: specifications and simulation," Computational Statistics & Data Analysis, vol. 42, pp. 513-533, 2003.

[5] W. R. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," Economic Geography, vol. 46, pp. 234-240, 1970.

[6] C. Brunsdon, et al., "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," Geographical Analysis, vol. 28, pp. 281-298, 1996.

[7] G. H. Pirie, "Distance," in International Encyclopedia of Human Geography, R. Kitchin and N. Thrift, Eds., ed Oxford: Elsevier, 2009, pp. 242-251.

[8] M. Charlton, et al., "Geographically Weighted Regression: Software for GWR," ed: National Centre for Geocomputation, 2007.

[9] M. F. Worboys, "Metrics and topologies for geographic space," in Advances in Geographic Information Systems Research II: Proceedings of the Symposium on Spatial Data Handling, 1996.

[10] B. Lu, et al., "Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data," Procedia Environmental Sciences, vol. 7, pp. 92-97, 2011.

[11] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR, 1973, pp. 267-281.

[12] S. Kullback and R. A. Leibler, "On Information and Sufficiency," The Annals of Mathematical Statistics, vol. 22, pp. 79-86, 1951.

[13] [C. M. Hurvich, et al., "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 60, pp. 271-293, 1998.

[14] P. Harris, et al., "Geographically weighted principal components analysis," International Journal of Geographical Information Science, vol. iFirst, pp. 1-20, 2011.