Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 7 (2011) 92–97

Procedia Environmental Sciences

1st Conference on Spatial Statistics 2011

Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data

Binbin Lu[a], Martin Charlton[a], A. Stewart Fotheringham[a] a*

[a]*National Centre for Geocomputation, NUI, Maynooth, Co.Kildare, Ireland*

Abstract

Geographically Weighted Regression (GWR) is a local modelling technique to estimate regression models with spatially varying relationships. Generally, the Euclidean distance is the default metric for calibrating a GWR model in previous research and applications; however, it may not always be the most reasonable choice due to a partition by some natural or man-made features. Thus, we attempt to use a non-Euclidean distance metric in GWR. In this study, a GWR model is established to explore spatially varying relationships between house price and floor area with sampled house prices in London. To calibrate this GWR model, network distance is adopted. Compared with the other results from calibrations with Euclidean distance or adaptive kernels, the output using network distance with a fixed kernel makes a significant improvement, and the river Thames has a clear cut-off effect on the parameter estimations.

*Keywords:*Geographically Weighted Regression; Non-Euclidean distance; Network distance; House price data.

1. Introduction

In recent years, there has been an increasing interest in the application and development of local forms of spatial analysis methods that produce local and mappable results instead of one-size-fits-all results from traditional global methods[1, 2]. Among these techniques, Geographically Weighted Regression (GWR)[3, 4] is proposed as a kind of local technique to estimate regression models with spatially varying relationships. This technique has been applied broadly in economics[5-8], sociology [9-11], urban and regional analysis [12-15], ecology and environment [16-20], etc. All these applications indicate that GWR is reliable and preferable for exploring the spatial heterogeneity in processes. In a way coinciding with

* Binbin Lu. Tel.: +353-1-7086731; fax:+353-1-7086456.
E-mail address:binbin.lu.2009@nuim.ie.

doi:10.1016/j.proenv.2011.07.017

Tobler's first law of geography[21], GWR makes a point-wise calibration around each regression point using a 'bump of influence': around each regression point nearer observations have more influence in estimating the local set of parameters than observations farther away[4]. This principle makes the distance measurement to be an essential element in the calibration of a GWR model. Until now, the Euclidean distance (ED) (straight line distance) has generally been adopted as the default for GWR applications. However, it is reasonable to ask: is the ED always the best choice for calibrating a GWR model?

The degree of connectedness between two places may not be optimally represented by a straight line. For example, an area might be divided by a river or buildings and connected internally by bridges or roads; surface distance would be more reasonable in a mountainous area. This inapplicability of ED is highlighted especially in human activities related studies. Actually the notion of distance metric (also termed as proximity) is a fundamental concept in any comprehensive ontology of space[22], and it has become an important part concerned in applying spatial techniques. In Geostatistics, non-Euclidean distances are suggested in previous studies according to their spatial features [23-26]. In this paper, we are working with a London house price dataset, and the connectivity between houses will be calculated using network distance.

The reminder of the paper is organized as follows. In the next section, we describe the basic methodology of GWR. Afterward, the sampled house price data are reviewed and a GWR model to explore the spatially varying relationship between house price and floor area is introduced, and this model is calibrated using ED and ND respectively with fixed and adaptive kernels. Finally, conclusions are summarized and future work is outlined.

2. GWR methodology

Geographically Weighted Regression (GWR) evolved from traditional linear regression methods by permitting the relationships between variables to vary spatially. A basic GWR model for each regression point could be written as:

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

where y_i is the dependent variable at location *i*, x_{ik} is the value of the kth explanatory variable at location *i*, $\beta_0\left(u_i, v_i\right)$ is the intercept parameter at location *i*, $\beta_k\left(u_i, v_i\right)$ is the local regression coefficient for the *k*th explanatory variable at location *i*, $\left(u_i, v_i\right)$ is the coordinate of location *i*, ε_i is the random error at location *i*.

For a GWR model, a point-wise calibration is made for each regression location independently. In this process weighted least squares is used, and the matrix calculation for the estimated regression coefficients could be expressed as:

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

where *X* is the matrix of the explanatory variables with a column of 1s for the intercept, *y* is the vector of the dependent variables, $\beta\left(u_i, v_i\right) = \left(\beta_0\left(u_i, v_i\right), \cdots, \beta_n\left(u_i, v_i\right)\right)$ is the vector of *n+1* local regression coefficients, $W\left(u_i, v_i\right)$ is the diagonal matrix denoting the geographical weighting of each observed data for regression point *i*.

Here, the weighting scheme $W\left(u_i, v_i\right)$ is based on the proximity of regression point *i* to the data points around *i*, and calculated from a kernel function. In practice, a common continuous choice is the Gaussian kernel function:

$$w_{ij} = \begin{cases} \exp[-\frac{1}{2}(\frac{d_{ij}}{b})^2] \; if \; d_{ij} < b \\ 0 \qquad otherwise \end{cases} \qquad (3)$$

where d_{ij} is the distance between regression point i and data point j, b is the distance threshold, also known as bandwidth. This is the key to apply a different distance metric and the simplest way is to substitute the d_{ij} in accordance with a kind of distance metric. In the meanwhile, a proper bandwidth b should also be selected based on the applying distance metric.

3. Data and GWR modelling

In this paper we use house price data for a sample of 372 properties sold within London during 2001. The study area is divided by the river Thames and this makes the ND measurements significantly different from ED, which is affected by the density of bridges along the river. The road network data used here is downloaded from OpenStreetMap[27]. A preview of used data is shown in figure 1(a).

As for house price data, a common quantity will be the average price per square meter (£/m^2) of a property. In this sense, we propose a regression model between the price and floor area, and its GWR expression could be written as:

$$P_i = \beta_0\left(u_i, v_i\right) + \beta_1\left(u_i, v_i\right) FLRAREA_i \tag{4}$$

where P_i is the sold price of a property in pound; *FLRAREA* is the floor area of the house in square meter. In order to facilitate the analysis of results, the locations of sampled houses are also used as regression points.

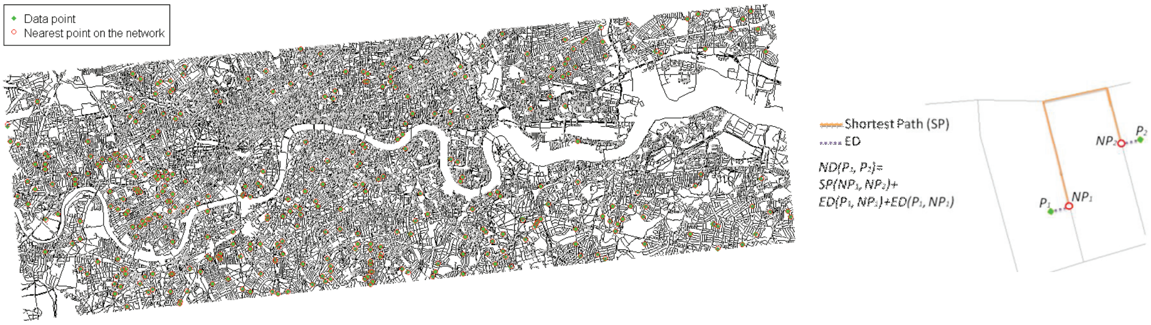

Figure 1 (a) Data points and corresponding nearest points on the given network (b) Illustration of computing ND

The first problem is how to calculate the network distances between regression points and observation points. In previous research, the ND between two points always refers to length of the shortest path between them on a given network. However, the sampled points are not always on a network, especially in this study (seen in figure 1). The usual way to overcome this is to find the nearest point or node on the network for substituting them. Consequently, on the context of locating each point with the nearest point, the ND between a pair of points consists of two parts: the length of shortest path between corresponding nearest points; the ED between them and their nearest points. As illustrated in figure 1, the nearest point corresponding to each house is located, and the ND could be computed following the method.

We calibrate the above model in five different ways: solve it as an ordinary least square (OLS) model; treat it as a GWR model and do the calibrations using ED and ND respectively with fixed and adaptive spatial kernels. As for the bandwidth for each GWR calibration, the cross validation (CV) approach[28] is employed for their selections. As shown by the diagnostic information in table 1, above all the GWR approaches make a much better performance than OLS; moreover, there is a significant improvement when the calibration is done with a fixed kernel and ND according to the smallest AIC value and largest Adjusted R square value, although the values are slightly smaller across all four GWR results.

Table 1 Comparison of Akaike Information Criterion (AIC) and adjusted R square values for different calibrations of the model

	OLS model	GWR model (Fixed & ED)	GWR model (Adaptive & ED)	GWR model (Fixed & ND)	GWR model (Adaptive & ND)
AIC	9529.11	9065.089	9078.045	9057.178	9078.187
Adjusted R square	0.4671916	0.8241892	0.8180545	0.8254611	0.8155602

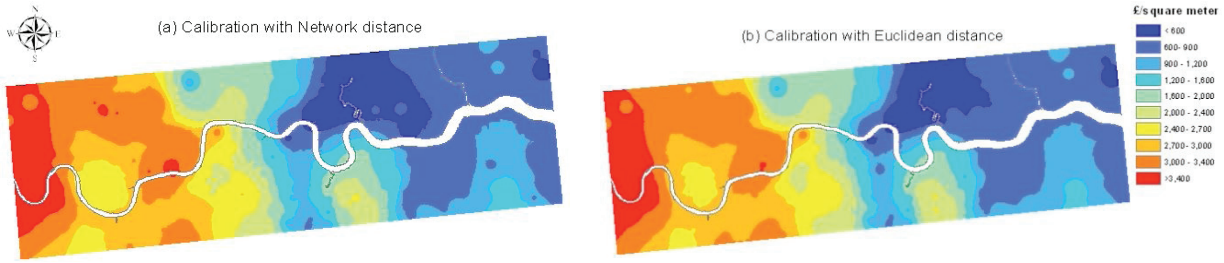

Figure 2 Maps of parameter β_1 estimation using ND and ED in fixed kernels

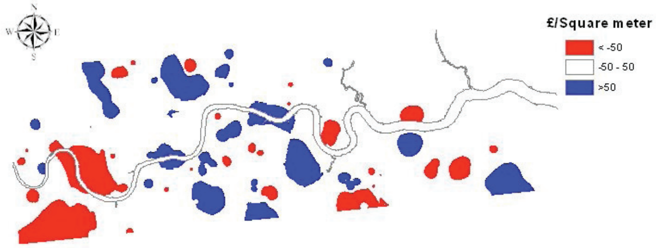

Figure 3 Residual map between parameter β1 estimation with ND and the one with ED

For a detailed comparison, we produce two maps of the parameter estimates using ND and ED with fixed kernels, shown in figure 2. In general, they are similar with each other and the estimates have almost the same spatial pattern in this study area. However, the result from ED seems smoother than the one from ND, in which there are many prominent spots. As shown in figure 3, we also produce a residual map by subtracting figure 2(b) from figure 2(a). This map shows that the change of estimations occurs in many blocks, which are immediately related with the road density in that area. Simultaneously the expected cut-off effect of the river Thames is clearly displayed, as almost half of significant estimation changes have taken place along it, especially within Fulham.

4. Conclusion and future work

This paper has used ND in calibrating a GWR model, and the results confirmed the feasibility of improving the performance of GWR via using ND instead of ED. The appearance of cut-off effect also indicates that the reachability distances between houses have a significant impact on the parameter estimation. However, the improvement in this case study is not huge, especially from the aspect of adjusted R square values. This may be due to the complexity of roads. In this case, we arbitrarily weighted the road segments with their physical length no matter what type they are, and calculated the

ND on this basis. Using the classification of roads or even the calculation of travel time might be better choices for this application.

Actually, the distance metric is somehow depended on a particular feature attribute. As a result, a different distance metric could be specified for each parameter in a multiple GWR model according to their properties. This forms an interesting future challenge.

Acknowledgements

Research presented in this paper was jointly funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan and China Scholarship Council (CSC). The authors gratefully acknowledge their support.

References

[1] A.S. Fotheringham, C. Brunsdon, Local Forms of Spatial Analysis, Geographical Analysis, 31 (1999) 340-358.
[2] A. Páez, Local Analysis of Spatial Relationships: A Comparison of GWR and the Expansion Method, in: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.) Computational Science and Its Applications – ICCSA 2005, Springer Berlin / Heidelberg, 2005, pp. 631-637.
[3] C. Brunsdon, A.S. Fotheringham, M.E. Charlton, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity, Geographical Analysis, 28 (1996) 281-298.
[4] A.S. Fotheringham, M.E. Charlton, C. Brunsdon, Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis, Environment and Planning A, 30 (1998) 1905-1927.
[5] T. Benson, J. Chamberlin, I. Rhinehart, An investigation of the spatial determinants of the local prevalence of poverty in rural Malawi, Food Policy, 30 (2005) 532-550.
[6] B. Huang, B. Wu, M. Barry, Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices, International Journal of Geographical Information Science, 24 (2010) 383-401.
[7] K. Yan, T. Marius, R. François Des, Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles, Journal of Geographical Systems, 8 (2006) 61-96.
[8] D. Yu, Modeling Owner-Occupied Single-Family House Values in the City of Milwaukee: A Geographically Weighted Regression Approach, GIScience & Remote Sensing, 44 (2007) 267-282.
[9] M. Cahill, G. Mulligan, Using Geographically Weighted Regression to Explore Local Crime Patterns, Social Science Computer Review, 25 (2007) 174-193.
[10] L. Waller, L. Zhu, C. Gotway, D. Gorman, P. Gruenewald, Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models, Stochastic Environmental Research and Risk Assessment, 21 (2007) 573-588.
[11] D. Wheeler, L. Waller, Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests, Journal of Geographical Systems, 11 (2009) 1-22.
[12] K. Ali, M.D. Partridge, M.R. Olfert, Can Geographically Weighted Regressions Improve Regional Analysis and Policy Making?, International Regional Science Review, 30 (2007).

[13] D.P. Mark, S.R. Dan, A. Kamar, M.R. Olfert, The Geographic Diversity of U.S. Nonmetropolitan Growth Dynamics: A Geographically Weighted Regression Approach, Land Economics, 84 (2008) 241-266.
[14] J. Gao, S. Li, Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using Geographically Weighted Regression, Applied Geography, 31 (2011) 292-302.
[15] J. Tu, Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression, Applied Geography, 31 (2011) 376-392.
[16] Y. Kamarianakis, H. Feidas, G. Kokolatos, N. Chrysoulakis, V. Karatzias, Evaluating remotely sensed rainfall estimates using nonlinear mixed models and geographically weighted regression, Environmental Modelling & Software, 23 (2008) 1438-1447.
[17] L.J. Young, C.A. Gotway, J. Yang, G. Kearney, C. DuClos, Linking health and environmental data in geographical analysis: It's so much more than centroids, Spatial and Spatio-temporal Epidemiology, 1 (2009) 73-84.
[18] S. Li, Z. Zhao, X. Miaomiao, Y. Wang, Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression, Environ. Model. Softw., 25 (2010) 1789-1800.
[19] M.J.S. Windle, G.A. Rose, R. Devillers, M.-J. Fortin, Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic, ICES Journal of Marine Science: Journal du Conseil, 67 (2010) 145-154.
[20] A. Gilbert, J. Chakraborty, Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida, Social Science Research, 40 (2011) 273-286.
[21] W.R. Tobler, A Computer Movie Simulating Urban Growth in the Detroit Region, Economic Geography, 46 (1970) 234-240.
[22] M.F. Worboys, Nearness relations in environmental space, International Journal of Geographical Information Science, 15 (2001) 633 - 651.
[23] F. Curriero, On the Use of Non-Euclidean Distance Measures in Geostatistics, Mathematical Geology, 38 (2006) 907-926.
[24] J. Grazzini, P. Soille, C. Bielski, On the use of geodesic distances for spatial interpolation, in: Proceedings of the 9th International Conference on GeoComputation, GeoComputation, Maynooth, Ireland, 2007.
[25] C.A.G. Crawford, L.J. Young, Geostatistics: What's Hot, What's Not, and Other Food for Thought, in: Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, P. R. China, 2008, pp. 8-16.
[26] E.S. Money, G.P. Carter, M.L. Serre, Modern Space/Time Geostatistics Using River Distances: Data Integration of Turbidity and E. coli Measurements to Assess Fecal Contamination Along the Raritan River in New Jersey, Environmental Science & Technology, 43 (2009) 3736-3742.
[27] OpenStreetMap community, United Kingdom OSM Highway, in: http://downloads.cloudmade.com (Ed.), 2010.
[28] A.S. Fotheringham, C. Brunsdon, M. Charlton, Geographically Weighted Regression: the analysis of spatially varying relationships, Wiley, 2002.