Semiparametric geographically weighted generalised linear modelling in GWR 4.0

T. Nakaya[1], A. S. Fotheringham[2], M. Charlton[2], C. Brunsdon[3]

[1]Department of Geography, Ritsumeikan University,
56-1 Tojiin-kita-machi, Kita-ku, Kyoto 603-8577, Japan
Telephone: (+81)-(0)75-467-8801
Fax: (+81)-(0)75-465-8296
Email: nakaya@lt.ritsumei.ac.jp

[2]National Centre for Geocomputation, John Hume Building
National University of Ireland, Maynooth, Co Kildare, Ireland
Telephone: (+353)-(0)-1-708-6455
Fax: (+353)-(0)-1-708-6456
Email: Stewart.Fotheringham@nuim.ie, Martin.Charlton@nuim.ie

[3]Department of Geography, Leicester University,
Leicester, LE1 7RH, UK
Telephone: (+44)-(0)-116-252-3843
Fax: (+44)-(0)-116-252-3854
Email: cb179@le.ac.uk

1. Introduction

The aim of this paper is to propose a generalised framework for semiparametric geographically weighted regression (S-GWR) by combining several theoretical aspects of geographically weighted regression (GWR). In this framework, we can implement model selection in order to judge which explanatory effects on the response variable are globally fixed or geographically varying in generalised linear modelling (GLM). This framework is implemented in a new version of the GWR software (GWR 4.0) which is soon to be released and which will be described.

To date, numerous theoretical and applied studies of GWR have been reported after the first seminal papers of GWR appeared (Fotheringham et al., 1996; Brunsdon et al, 1996). Here, we focus on two important extensions of GWR; geographically weighted generalised linear modelling (GWGLM) and semiparametric extension of GWR. While the original GWR assumes that the response is a continuous variable and the error term follows a Gaussian (normal) distribution, GWGLM enables us to fit generalised linear models with geographically local coefficients to accommodate commonly encountered types of response including count and binomial variables with likelihood functions of non-normal errors.

Although Fotheringham et al. (2002) described a geographically local scoring algorithm to estimate geographically local coefficients of GWGLM, issues of inference about estimated coefficients were generally ignored. Nakaya et al. (2005) derived standard errors of coefficients and degrees-of-freedom for model selection indicators by focusing on geographically weighted Poisson regression (GWPR) and its semiparametric extension. Here, we generalize semiparametric GWPR to S-GWR and associate it with model diagnostics to assess the geographical variability of coefficients. The new software, GWR 4.0, provides a user-friendly platform to calibrate S-GWR models allowing the user to experiment with which coefficients are spatially varying and which are fixed.

2. S-GWR models

Geographically varying coefficient models are defined in the generalized linear model (GLM) framework. Suppose we define a linear predictor with geographically varying coefficients as,

$$\eta_i = \sum_k \beta_k(u_{i,}, v_i) x_{k,i} = \mathbf{x}_i^t \boldsymbol{\beta}(\mathbf{u}_i) \ ,$$

where $x_{k,i}$ and β_k is the *k*th explanatory (independent) variable and its coefficient, respectively. In this model, the coefficients vary depending on the geographical coordinate of the location, $\mathbf{u}_i = (u_{i,}, v_i)$. The expected value of response of the *i*th observation, $E[y_i]$, is related to the linear predictor via a link function, *g*;

$$g(E[y_i]) = \eta_i$$

The log-likelihood of the observation is defined by a distributional function of the exponential family with the canonical, η_i, and dispersion parameters, φ_i as well as three functional components of exponential family, *a*, *b* and *c*;

$$\log f(y_i \mid \eta_i, \varphi_i) = \frac{y_i \eta_i - b(\eta_i)}{a(\varphi_i)} + c(y_i, \varphi_i) \ .$$

This framework covers commonly used regression models including Gaussian, Poisson and logistic variants of GWR.

GWR (Gaussian): $y_i \sim N\left[\eta_i, \sigma^2\right]$
GWPR (Poisson): $y_i \sim \text{Poisson}\left[\exp(\eta_i)\right]$
GWLR (Logistic): $y_i \sim \text{Bernoulli}\left[\text{logistic}(\eta_i)\right]$

GWGLM is a method to estimate a vector of local coefficients focusing on the *i*th regression point by solving the following maximisation problem of the geographically weighted log-likelihood of the model,

$$\hat{\boldsymbol{\beta}}(\mathbf{u}_i) = \arg\max \sum_j^n \left\{ \log f(y_j \mid \hat{\eta}_{(i)j}, \hat{\varphi}_{(i)j}) \cdot w_{ij}\left(\left\|\mathbf{u}_i - \mathbf{u}_j\right\|\right) \right\}$$

where the hat symbol means prediction;

$$\hat{\eta}_{(i)j} = \eta_j\left(\hat{\boldsymbol{\beta}}(\mathbf{u}_i)\right), \ \hat{\varphi}_{(i)j} = \varphi_j\left(\hat{\boldsymbol{\beta}}(\mathbf{u}_i)\right) \ .$$

These two working variables are fitted canonical and dispersion parameters for the prediction of the response at the *j*th location with coefficients at the *i*th regression point. The geographical weight of the *j*th observation at the *i*th regression point, w_{ij}, is introduced here as a non-negative and monotonously decreasing function of the distance between the regression point *i* and the *j*th observation location, such as a Gaussian kernel function:

$$w_{ij} = \exp\left(-\frac{1}{2}\frac{\left\|\mathbf{u}_i - \mathbf{u}_j\right\|}{G}\right),$$

where the parameter *G* (called the *bandwidth*) regulates the kernel size.

S-GWR as semiparametric GWGLM includes partially linear terms of explanatory effects on the response in the canonical parameter;

$$\eta_i = \sum_k \beta_k(u_{i,}, v_i) x_{k,i} + \sum_l \gamma_l z_{l,i}$$

where $z_{l,i}$ is lth explanatory variable and γ_l is its coefficient that is constant over space. Combining geographically local scoring and back-fitting algorithms, we can compute the estimates of coefficients and indicators for model diagnostics including standard errors of coefficients, degree-of-freedom and information criterion such as AICc (corrected AIC) to decide an optimal bandwidth size and model comparisons (cf. Nakaya et al., 2005).

3. Model selection of S-GWR

3.1 Assessment of geographical variability of coefficient

An advantage of S-GWR is that we can incorporate a fixed effect of a subset of explanatory variables on the response variable due to prior knowledge. However, it is not always obvious which coefficients should be assumed to be fixed or varying. A natural way to overcome this difficulty is to conduct empirical model comparisons of different semiparametric models having different combinations of fixed and varying coefficients. Mei et al. (2004; 2006) proposed F test schemes for this kind of model selection. However, considering the situation that an optimal bandwidth size of a geographical kernel for fitting GWR is normally carried out by a model selection indicator, it would be more appropriate to conduct such model comparisons for assessing geographical variability of coefficients of GWR by using a model selection indicator.

To assess the variability of the kth coefficient, we can compare two models; a fitted GWGLM model (pivot model) and a model in which only the kth coefficient is switched to be constant while the other coefficients vary spatially. If the pivot model is better than the model with the kth coefficient fixed, as judged by a model comparison criterion such as AICc, we can claim that the kth coefficient varies spatially. The test routine in GWR 4.0 repeats this comparison for each relationship in the model.

3.2 GtoF / FtoG automated variable selection

GWR 4.0 contains two separate fitting techniques for automated variable selection of S-GWR models. One is the GtoF (from geographically varying to fixed) variable selection routine which executes a series of model comparisons to search for an optimal combination of varying and fixed term given explanatory and response variables. The concept is similar to that of step-wise variable selection. Firstly, a model comparison is repeated between the originally fitted GWR model and a model in which only one coefficient is switched to be constant while the others remain spatially varying. The optimal model is now selected from the original GWR model and a set of models in which one parameter is fixed. If this optimal model is the full GWR model, the process stops. If the optimal model contains a fixed effect, a new set of model comparisons is then made by making each of the remaining spatially varying coefficients constant in turn and repeating this procedure until no further improvement in model fit can be obtained by making a coefficient constant instead of spatially varying.

An alternative model selection procedure also implemented in GWR 4.0 is FtoG (from fixed to geographically varying) which is the reverse procedure to that described above. In this case, the default model is the global one and the first round of model comparisons is made by allowing each parameter in turn to be spatially varying. Model selection is made in the same manner by selecting an optimal model based on AICc and then allowing a second round of parameters to be spatially varying etc until no model improvement is possible.

4. GWR 4.0

GWR 4.0 is a new release of the current GWR 3.0 software. Compared to the previous version, the user interface is largely rebuilt to use tabbed sub-windows so that a modelling session intuitively proceeds in a step-by-step manner (Figure 1). Also, a wider range of options related to GWGLM including geographically variability assessment and automated variable selection routines explained above are available. It is executable under a MS-Windows environment with .Net Framework 3.5. The software will be demonstrated in this presentation.

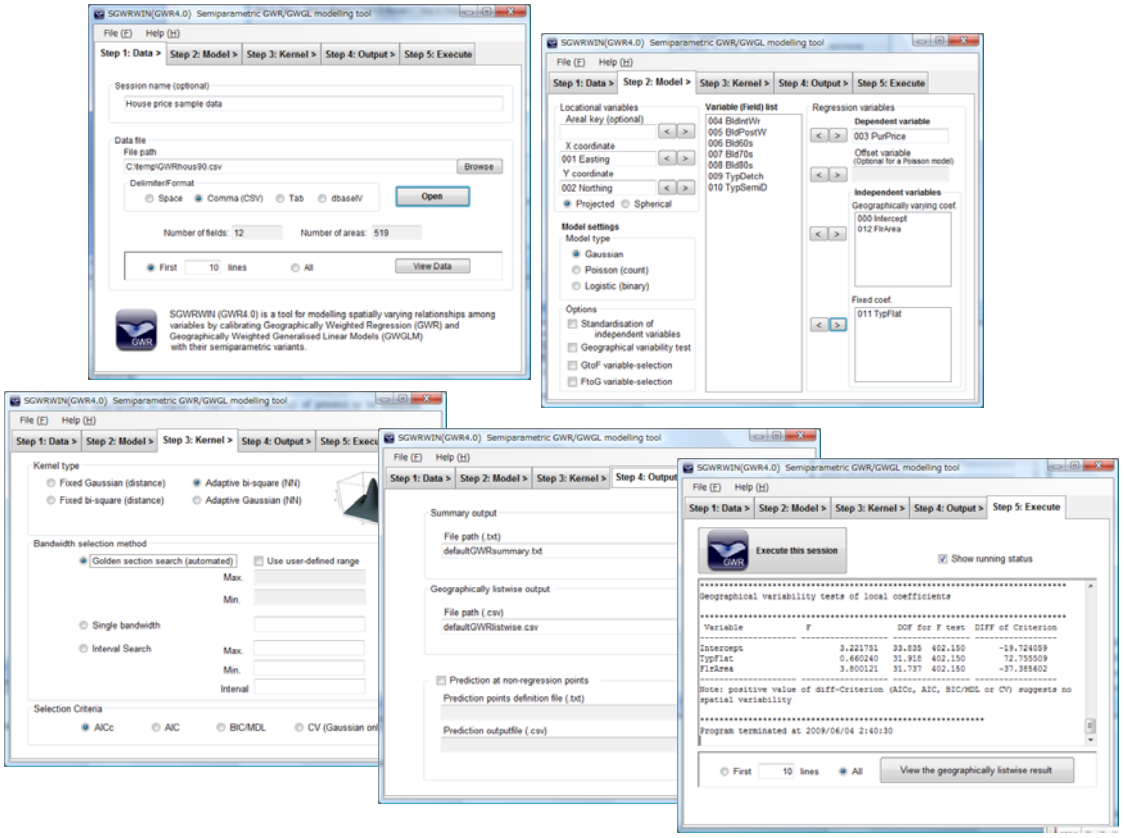

Figure 1. Screenshots of GWR4.0 interface.

5. Conclusion

In this paper, we describe GLM-based semiparametric geographically weighted regression (S-GWR) which allow mixing geographically varying and fixed coefficients in a generalised linear model. It is also possible to explore which explanatory terms should be varying or fixed, through model comparisons between possible different S-GWR models. GWR 4.0 has been developed as a platform for the practical implementation of S-GWR modelling with new methods of geographical variability assessments for estimated coefficients and automated model selection to search for an optimal combination of fixed and varying explanatory terms in a model.

References

Brunsdon C, Fotheringham AS and Charlton M, 1996, Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28, 281-289.

Fotheringham AS, Charlton M and Brunsdon C, 1996 The geography of parametrer space: an investigation into spatial nonstationarity, *International Journal of GIS* , 10, 605-627

Fotheringham AS, Brunsdon C and Charlton M, 2002, *Geographically weighted regression*. Wiley, Sussex, UK.

Mei C-L, He S-Y and Fang K-T, 2004, A note on the mixed geographically weighted regression model. *Journal of Regional Science*, 44, 143-157.

Mei C-L, Wang N and Zhang W-X, 2006, Testing the importance of the explanatory variables in a mixed geographically weighted regression model. *Environment and Planning A*, 38, 587-598.

Nakaya T, Fotheringham AS, Charlton M and Brunsdon C, 2005, Geographically weighted Poisson regression for disease associative mapping, *Statistics in Medicine*, 24, 2695-2717.