Geographically Weighted Regression Models for Ordinal Categorical Response Variables—Applications to Air Pollution and Quality of Life in Beijing, China

Guanpeng Dong*[1] and Alex Singleton[2]

[1]Department of Geography and Planning, University of Liverpool

[2]Department of Geography and Planning, University of Liverpool

*Email: guanpeng.dong@liverpool.ac.uk

Abstract

Ordinal categorical responses are commonly seen in geo-referenced survey data while spatial statistics tools for modelling such type of outcomes are rather limited. The paper extends the local spatial modelling framework to accommodate ordinal categorical response variables by developing a Geographically Weighted Ordinal Regression model. The GWOR model offers a proper statistical tool to analyse spatial data with ordinal categorical responses, allowing for the exploration of spatially varying relationships in the model. Based on a geo-referened life satisfaction survey data in Beijing, China, the proposed methodology is employed to explore the socio-spatial variation of life satisfaction and how environmental quality is associated with life satisfaction. We find that air pollution is negatively associated with life satisfaction, which is both statistically significant and spatially varying. The economic valuation of air pollution results show that residents are willing to pay about 2.6% of their annual income for per unit air pollution abatement, on average.

Keywords: GWR, ordinal response variables, air pollution

1 Introduction

This paper contributes to the ongoing methodological development of GWR models by proposing a geographically weighted ordinal response regression model (GWOR) in order to properly model spatial data sets with ordinal categorical responses. Ordinal response variables are commonly used in social science research, especially when the focus is in relation to individual opinions and attitudes towards events, or subjective assessment of life experiences such as satisfaction and happiness. Detailed descriptions of the application scope of ordinal response variables in a variety of disciplines are provided in Agresti (2010) and Greene and Hensher (2010). The motivation of developing a GWOR model lies in our interest of exploring the socio-spatial variation of life satisfaction in Beijing, China and examining the role of geography in how life satisfaction are linked to air pollution and socio-economic status.

*

To date, GWR models have been developed most often assuming the outcome variable under investigation distributed as a Gaussian process, with few notable exceptions in Nakaya et al. (2005) where a geographically weighted Poisson regression model has been developed for exploring disease outcomes following a Poisson distribution. This paper fills the gap by developing a GWOR model for ordinal categorical response variables, offering a tool for the exploration of spatially varying relationships between an ordinal response variable and predictor variables. Model estimation draws upon the local likelihood approach and is implemented through iterative numerical optimisation approaches (detailed below). The model allows for great flexibility in terms of model specification, including different link functions (logit and probit) for the cumulative distribution of the categorical responses and a mixed model specification (also termed as semi-parametric models in Nakaya et al. (2005)) in which regression coefficients of some variables are spatially varying while coefficients of other variables are kept spatially invariant.

2 A non-spatial ordinal response model

Following Agresti (2010) and Greene and Hensher (2010), we use a latent variable approach to formulate an ordinal categorical response regression model due to its intuitive link to the linear regression models familiar to most quantitative researchers.[1] Denote Y_i^* as a latent continuous outcome variable and x_i a set of predictor variables such as income, air pollution, and others. A linear regression model links Y_i^* to x_i,

$$Y_i^* = x_i\boldsymbol{\beta} + \epsilon_i; \forall i \in N \tag{1}$$

where i indexes each observation and N, the sample size. $x_i = [x_{i,1}, x_{i,2}, \dots, x_{i,p}]$ is a row-vector of predictor variable values for observation i while $\boldsymbol{\beta}$ is a column-vector of regression coefficients to estimate. The mapping of the unobservable Y_i^* to the observed categorical response Y_i depends on a set of cut-off points or threshold values $[\alpha_0, \alpha_1, \dots, \alpha_J]$ on the scale of Y_i^*,

$$\begin{aligned} Y_i &= 1 \qquad if \quad \alpha_0 < Y_i^* \leq \alpha_1 \\ &= j \qquad if \quad \alpha_{j-1} < Y_i^* \leq \alpha_j \qquad j = 2, \dots, J-1 \\ &= J \qquad if \quad \alpha_{J-1} < Y_i^* \leq \alpha_J \end{aligned} \tag{2}$$

where $j = 1, \dots, J$ indicates the category that each response falls into. Ordinal response models focus on the cumulative probability of an observation falling in category j or below, which is expressed as,

$$P(Y_i \leq j) = P(Y_i^* \leq \alpha_j) = P(\epsilon_i \leq \alpha_j - x_i\boldsymbol{\beta}) = F(\alpha_j - x_i\boldsymbol{\beta}) \tag{3}$$

Clearly, based on the different specifications of density functions for ϵ, the cumulative probability of $P(Y_i \leq j)$ has different forms: $\frac{1}{1+e^{-(\alpha_j - x_i\boldsymbol{\beta})}}$ if a logistic density was specified for ϵ, and $\Phi(\alpha_j - x_i\boldsymbol{\beta})$ if a Normal density was specified. The logistic specification was slighted favoured due to its simplicity in terms of model parameter interpretation over Normal distributions (Agresti, 2010). The probability

[1] The cumulative link (logit or probit) function approach is another popular way to formulate the ordinal response regression models. The two different approaches lead to the same model specification (Agresti, 2010). The latent variable approach is predominantly used in the environmental valuation literature, so we follow this convention.

of $Y_i = j$ conditioning on x_i is given by: $F[Y_i^* \leq \alpha_j] - F[Y_i^* \leq \alpha_{j-1}]$. While the GWOR models developed here allows for both Normal and logistic densities for ϵ, the interpretation is demonstrated by using the latter.

3 Developing a geographically weighted ordinal response model

We now introduce the geographically weighted ordinal response models that allows for regression coefficients varying across space. Following the GWR notational convention (Fotheringham et al., 2003), let $u_i(u_{i,x}, u_{i,y})$ be the geographical coordinates, describing the location, say residence address, of observation i. Equations (1) and (2) can be rewritten as (note the model intercept term is dropped),

$$\begin{aligned} Y_i^* &= x_i\boldsymbol{\beta}(u_i) + \epsilon_i; \forall i \in N \\ Y_i &= 1 \quad if \quad \alpha_0(u_i) < Y_i^* \leq \alpha_1(u_i) \\ &= j \quad if \quad \alpha_{j-1}(u_i) < Y_i^* \leq \alpha_j(u_i) \qquad j = 2, \ldots, J-1 \\ &= J \quad if \quad \alpha_{J-1}(u_i) < Y_i^* \leq \alpha_J(u_i) \end{aligned} \tag{4}$$

Both the $J - 1$ cut points and the coefficient vector $\boldsymbol{\beta}$ to estimate are associated with the location indicator u_i, relaxing the restrictive assumption of spatially invariant regression coefficients. An additional flexibility of the GWOR over non-spatial ordinal response model is to allow the locational heterogeneity in how the latent variable space is divided into J categories. Following (Páez et al., 2002; Nakaya et al., 2005; Loader, 2006), a geographically weighted local likelihood approach was employed to estimate GWOR parameters.

3.1 Model estimation

The log-likelihood function at the fit (or focal) location i is expressed as,

$$l(\boldsymbol{\theta}(u_i)) = \sum_{k=1}^{N}\sum_{j=1}^{J} w_{ik}y_{kj}Log[F(\alpha_j(u_i) - x_k\boldsymbol{\beta}(u_i)) - F(\alpha_{j-1}(u_i) - x_k\boldsymbol{\beta}(u_i))] \tag{5}$$

where w_{ik} is the geographical weight placed on the kth observation when fitting the local regression model at location i. A weighting function is needed to determine the rate of decrease in weights assigned to observations when moving further away from the fit point. A commonly employed weighting kernel is the Gaussian kernel, in which $w_{ik} = e^{-0.5\frac{(|u_i - u_k|)}{D}}$ with $|u_i - u_k| = (u_{i,x} - u_{k,x})^2 + (u_{i,y} - u_{k,y})^2$. The parameter D is the kernel bandwidth, which controls how fast the weights decline with increasing distance from the fit or focal location. Larger bandwidth implies a slower weights decay. Alternative kernel functions such as bi-square kernel are also widely used (Páez et al., 2002; Fotheringham et al., 2003). A further decision is to choose between a fixed kernel approach where bandwidth is fixed for all fit locations, and an adaptive kernel approach in which bandwidth varies with each fit location while keeping the number of nearest observations from each fit point same across a study area.

Again, an iterative maximisation algorithm is needed to estimate model parameters at location i. The maximum of each local log-likelihood function (Equation (7)) can be found by using the Newton-Raphson approach, as illustrated in the following equation (Franses and Paap, 2001, p.119),

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} - H(\boldsymbol{\theta}_{m-1})^{-1} G(\boldsymbol{\theta}_{m-1}) \tag{6}$$

where $G(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ are the first-order and second-order derivatives of the local Log-likelihood function $l(\boldsymbol{\theta}(u_i))$ with respect to model parameters $\boldsymbol{\theta}$, which are also known as the gradient and Hessian matrix of a likelihood function, respectively. Based on Equation (7), the gradient vector can be obtained as,

$$G(\boldsymbol{\theta}(u_i)) = \frac{\partial l(\boldsymbol{\theta}(u_i))}{\partial \boldsymbol{\theta}(u_i)} = \sum_{k=1}^{N} \sum_{j=1}^{J} \Big[\frac{w_{ik} y_{kj}}{P(Y_k = j|x_k)} \frac{\partial P(Y_k = j|x_k)}{\partial \boldsymbol{\theta}(u_i)} \Big] \tag{7}$$

It is clear that only the conditional probability $P(Y_k = j|x_k)$ is involved in deriving the gradient vector. For notational simplicity, we denote $P(Y_k = j|x_k)$ as $P(Y_k = j)$ and the first-order derivative of the cumulative probability function $F(.)$ as $f(.)$. Then we have,

$$\begin{aligned} \frac{\partial P(Y_k = j)}{\partial \boldsymbol{\theta}(u_i)} &= \Big[\frac{\partial P(Y_k = j)}{\partial \boldsymbol{\beta}(u_i)}, \frac{\partial P(Y_k = j)}{\partial \alpha_1(u_i)}, \ldots, \frac{\partial P(Y_k = j)}{\partial \alpha_{J-1}(u_i)} \Big]' \\ &= [x_k(f(\alpha_{j-1}(u_i) - x_k\boldsymbol{\beta}(u_i)) - f(\alpha_j(u_i) - x_k\boldsymbol{\beta}(u_i))), \\ &\quad 0, \ldots, -f(\alpha_{j-1}(u_i) - x_k\boldsymbol{\beta}(u_i)), f(\alpha_j(u_i) - x_k\boldsymbol{\beta}(u_i)), \ldots, 0]' \end{aligned} \tag{8}$$

a $(p + J - 1)$ column vector. The Hessian follows naturally by further obtaining the derivative of Equation (9) with respect to $\boldsymbol{\theta}(u_i)'$.

4 Empirical analysis of life satisfaction and air pollution

4.1 Data and variables

Our analysis mainly draws on a life satisfaction survey data conducted in 2013 in Beijing, a key aim of which was to assess the socio-spatial variation in residents' life satisfaction. A key features of this survey data is its records of residence addresses for each respondent. Based on this information, the life satisfaction data have been geo-coded, which makes possible of extraction of local air pollution levels and other locational variables for each residence.

Life satisfaction is our outcome variable measured by asking this question: Overall, how are you satisfied with your life? The responses were quantified on a 5-point Likert scale ranging from 1 (very unsatisfied) to 5 (very satisfied). The majority ($>$ 60 percent) of respondents were satisfied with their lives while about 27.5 percent of respondents rated life satisfaction level as fair.

Air pollution data is compiled from the real-time air pollution monitoring data, hosted by the Beijing Environmental Protection Bureau (BJEPB). There are 35 air pollution monitoring sites in Beijing, producing hourly readings for various air pollutants. As the particulate matter with a diameter of $2.5\mu m$ or less ($\mathrm{PM}_{2.5}$) is of particular toxic and more strongly linked to health issues than other air pollutants in China (Zhang et al., 2013), we use the concentration of $\mathrm{PM}_{2.5}$, measured by unit of

$\mu g/m^3$, as a proxy variable for air pollution. Following Ferreira et al. (2013), an inverse distance weighted interpolation (IDW) approach was employed to extract the annual mean air pollution levels in 2013 for each residence location.

4.2 Model estimation results

Our final results are based on a fixed Gaussian kernel approach with a distance bandwidth of 15.8 kilometres obtained by using a cross-validation procedure. We also estimated an adaptive GWOR model with an optimum bandwidth of the nearest 956 neighbours, which accounts for roughly 36% of the total sample. The obtained spatial patterns of local coefficients are quite similar. A Monte Carlo permutation approach, outlined above, was used to test whether the spatial variations in local coefficients are statistically significant, or due to random sampling uncertainty. The non-stationary test results suggest that spatial heterogeneity in the associations of income, air pollution and locational factors to life satisfaction is unlikely due to random sampling uncertainties.

Looking at the distribution of the local coefficients of income (Log Income), there is relatively large variability, shown by the difference between the 2.5-th and 97.5-th percentiles of the distribution of local coefficients and also the large standard deviation). Figure 1 depicts the interpolated surface of the associations between income and life satisfaction with breaking points being the quintiles of local coefficients. Approximate significance inferences of local coefficients are also presented in the map. Each residence location is represented by a cross symbol and the statistical significance is indicated by different colours: black indicating significant and white insignificant at the conventional 5% significance level, depending on whether the absolute t-values of local coefficients are above 1.96 or not. From Figure 1, we see that the income-life satisfaction associations are increasing from the central urban area of Beijing to the suburban areas especially in the northwest and southwest Beijing.

The local coefficients of air pollution also exhibit large spatial variability, indicated by a roughly 41 per cent increase in the local coefficients from the lower quantile to the upper quantile of the distribution. The Monte Carlo non-stationary test for air pollution yields a p-value of 0.01, producing strong evidence on the spatial variation in how air pollution is related to life satisfaction. Figure 2 shows an interesting spatial pattern of local pollution coefficients. Overall, the central urban areas of Beijing see a stronger negative relationship between air pollution and life satisfaction than the suburban areas do, indicating a stronger aversion of air pollution for residents living in central urban areas.

5 Conclusion

In this paper, we extend the geographically weighted regression modelling framework to accommodate categorical response variables that are measured on an ordinal scale by developing a GWOR model. Ordinal response variables prevail in survey data and are increasingly explored in a wide range of social science disciplines, although in a way that the spatiality of the data under study is not usually taken into account seriously. The proposed methodology offers a flexible exploratory tool to explore the spatial aspects of data and address the issue of spatial heterogeneity. The usefulness of the proposed methodology is demonstrated by exploring the socio-spatial variation of resident's life satisfaction in urban Beijing and modelling the spatially varying relationships between life satisfaction and income, air pollution and other locational factors.

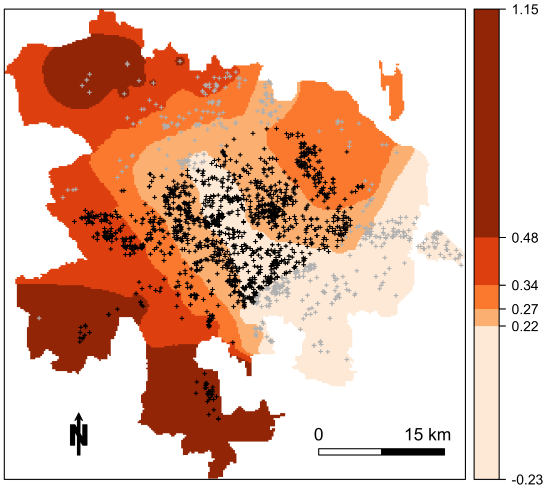

Figure 1: The spatial variation in the association between income and life satisfaction

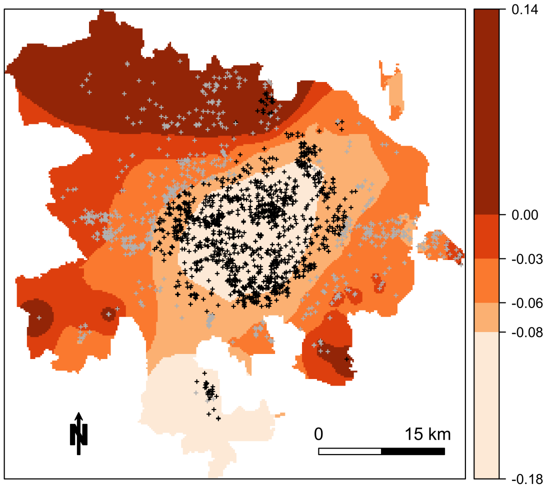

Figure 2: The spatial variation in the association between air pollution and life satisfaction

Agresti, A.
2010. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Ferreira, S., A. Akay, F. Brereton, J. Cuñado, P. Martinsson, M. Moro, and T. F. Ningal
2013. Life satisfaction and air quality in europe. *Ecological Economics*, 88:1–10.

Fotheringham, A. S., C. Brunsdon, and M. Charlton
2003. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Franses, P. H. and R. Paap
2001. *Quantitative models in marketing research*. Cambridge University Press.

Greene, W. H. and D. A. Hensher
2010. *Modeling ordered choices: A primer*. Cambridge University Press.

Loader, C.
2006. *Local regression and likelihood*. Springer Science & Business Media.

Nakaya, T., A. S. Fotheringham, C. Brunsdon, and M. Charlton
2005. Geographically weighted poisson regression for disease association mapping. *Statistics in medicine*, 24(17):2695–2717.

Páez, A., T. Uchida, and K. Miyamoto
2002. A general framework for estimation and inference of geographically weighted regression models: 1. location-specific kernel bandwidths and a test for locational heterogeneity. *Environment and Planning A*, 34(4):733–754.

Zhang, A., Q. Qi, L. Jiang, F. Zhou, and J. Wang
2013. Population exposure to pm 2.5 in the urban area of beijing. *PloS one*, 8(5):e63486.