ARTICLE IN PRESS

Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Geographically weighted regression models for ordinal categorical response variables: An application to geo-referenced life satisfaction data

Guanpeng Dong[a,*], Tomoki Nakaya[b], Chris Brunsdon[c]

[a] Department of Geography and Planning, University of Liverpool, Room 713, Roxby Building, Chatham St, Liverpool L69 7ZT, UK
[b] Department of Geography, Ritsumeikan University, Japan, 56-1 Tojiin-kita-machi, Kita-ku Kyoto, 603-8577, Japan
[c] The National Centre for Geocomputation, Maynooth University, Iontas Building, Maynooth University, Maynooth, Kildare, Ireland

ARTICLE INFO

Keywords:
Geographically weighted regression
Spatial heterogeneity
Ordinal response variables
Air pollution
Environmental valuation
Beijing

ABSTRACT

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

1. Introduction

Geographically weighted regression (GWR) has been established as a flexible framework for modelling spatially varying relationships between predictor variables and an outcome variable (Brunsdon, Fotheringham, & Charlton, 1996; Fotheringham, Brunsdon, & Charlton, 2003). Recent years have seen active methodological development of GWR models due to an increasing demand of applying localised spatial models to data with complex structures and non-Gaussian types of outcome variables. For instance, GWR models have been extended to explore spatiotemporal data by incorporating temporal correlations between an observation at time period *t* and spatially nearby observations at previous periods into the overall local weights matrix for estimation (Fotheringham, Crespo, & Yao, 2015; Huang, Wu, & Barry, 2010). Harris, Dong, and Zhang (2013) presented a contextualized GWR model, thereby the contextual similarities of observations (e.g. similarity in the attributes of neighbourhoods where houses are located, measured by certain distance metric) were incorporated into the local weights matrix for implementing each local regression model. The key idea underlying this line of GWR model extension is to achieve a better or more realistic representation of spatial relationships between observations. Other methodological elaborations of GWR include developing formal statistical tests of spatial heterogeneity (Leung, Mei, & Zhang, 2000), and the use of different distance metrics in constructing the spatial weights matrix (Lu, Charlton, Brunsdon, & Harris, 2016).

This paper contributes to the ongoing GWR developments by extending a geographically weighted ordinal regression model (GWOR) for properly exploring spatial data with ordinal categorical response variables. Ordinal response variables are commonly seen in social science research, especially when the research focus is in relation to individual opinions and attitudes towards events, or subjective assessment of life experiences such as life satisfaction and happiness. Detailed descriptions of the application scopes of ordinal response variables in a variety of social science disciplines are provided in Agresti (2010) and Greene and Hensher (2010). The motivation of extending a GWOR model lies in two aspects. The first is to address the issue of limited methodological options to deal with increasingly available geo-referenced survey data in the local spatial modelling literature. Such data usually quantify important information via categorical variables. Secondly, we are interested in exploring the socio-spatial variation of life satisfaction in Beijing, China and examining potential spatial heterogeneity in the association between life satisfaction and air pollution. The GWOR model, demonstrated by examining a geo-referenced life satisfaction data, can be applied to other spatial data.

Life satisfaction data are often collected based on surveys, in which questions such as "Overall, how satisfied are you with your life?" are

* Corresponding author.
E-mail addresses: Guanpeng.dong@liverpool.ac.uk (G. Dong), nakaya@ritsumei.ac.jp (T. Nakaya), Christopher.Brunsdon@nuim.ie (C. Brunsdon).

https://doi.org/10.1016/j.compenvurbsys.2018.01.012
Received 2 September 2017; Received in revised form 18 December 2017; Accepted 29 January 2018

asked (e.g. Welsch & Ferreira, 2014). Responses to the question are usually recorded on a Likert scale, ranging from one being very unsatisfied to five being very satisfied, for instance. Life satisfaction scores or ratings are quantitative in nature but the between-category distances are unknown—the distance between categories of one and two might be quite different from that between categories of four and five (Agresti, 2010). This differs substantially from a Gaussian variable for which per unit difference is comparable. It has been shown that applying a linear regression model to an ordinal categorical variable would cause issues to model estimation and statistical inferences for regression coefficients, likely producing misleading model results (Agresti, 2010).

To date, the development of GWR focuses on outcome variables following a Gaussian (or Normal) distribution, with few notable exceptions in Nakaya, Fotheringham, Brunsdon, and Charlton (2005) where a geographically weighted Poisson regression model has been developed for exploring disease outcomes, and in Fotheringham et al. (2003) for a geographically weighted Binomial regression models. Closely related to this study, McMillen and McDonald (2004) presented a preliminary extension of GWR to geo-referenced ordinal response variables by proposing a locally weighted ordered probit model. However, discussions on model specifications such as choices of different link functions for the cumulative probabilities of responses and choices of adaptive or fixed kernels, and on approaches to test statistical significance of spatial heterogeneity in regression coefficients are not provided.

This paper extends the work of McMillen and McDonald (2004) by offering flexible tools to explore the potential spatial variability in relationships between an ordinal response variable and predictor variables. The estimation of GWOR model draws upon the locally weighted likelihood approach via an iterative numerical optimisation procedure (detailed below). It allows for great flexibility in model specification, including different link functions (logit and probit) for the cumulative probabilities of the responses, and a mixed model specification in which regression coefficients of some variables are spatially varying while coefficients of other variables are kept spatially invariant (e.g. Mei, Xu, & Wang, 2016). The R code for implementing various GWOR models are provided in the *Supplementary Information* of the paper.

Based on a geo-referenced life satisfaction survey data in Beijing, this study explores how life satisfaction is spatially linked to air pollution and other factors. There has been a surge of using life satisfaction data to evaluate environmental amenities such as air quality, the economic value of which cannot be directly observed through market transactions (Ferreira & Moro, 2013; MacKerron & Mourato, 2009; Welsch, 2006). The theoretical underpinnings of life satisfaction based environmental evaluation approach are comprehensively reviewed in Welsch and Ferreira (2014). At its heart, the subjective life satisfaction is regarded as the experienced utility of individuals, and by estimating the life satisfaction equation with environmental quality indicators and income included, the (marginal) willingness to pay (WTP) for environmental quality can be estimated (e.g. Ferreira & Moro, 2010). The issue, however, is that the life satisfaction equation was predominantly estimated by using aspatial regression models, implicitly assuming WTP for environmental quality improvement to be constant across space. This is a rather restrictive assumption. It is likely that people living in different locations with varying socio-economic characteristics tend to have different preferences for air quality and thus varying WTP for air quality improvement or air pollution abatement (Bayer, Ferreira, & McMillan, 2007; Ferreira & Moro, 2013). The GWOR model enables us to estimate the spatially varying associations between life satisfaction and air pollution and income, taking locational heterogeneity into account.

The remainder of this paper is organised as follows. Section 2 provides an overview of non-spatial ordinal regression models. In Section 3, we describe the GWOR model and provide details of model estimation. Section 4 applies the GWOR model to explore the socio-spatial variation of life satisfaction and estimate the economic value of air quality in Beijing. Conclusions are provided in Section 5.

2. A non-spatial ordinal regression model

Following Agresti (2010) and Greene and Hensher (2010), we use a latent variable approach to formulate an ordinal regression model due to its intuitive link to the simple linear regression models. 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,

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

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 of 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^*: $Y_i = j$ if $a_{j-1} < Y_i^* \le a_j$, $j = 1, \dots, J$ where J are the number of response categories. Ordinal regression models focus on the cumulative probability of an observation falling in category j or below, which is expressed as,

$$\mathrm{P}(Y_i \le j) = \mathrm{P}(Y_i^* \le a_j) = \mathrm{P}(\epsilon_i \le a_j - x_i\boldsymbol{\beta}) = F(a_j - x_i\boldsymbol{\beta}). \tag{2}$$

Different specifications of the density function for ϵ leads to different forms of cumulative probabilities for $\mathrm{P}(Y_i \le j)$: $1/(1 + exp(-a_j + x_i\boldsymbol{\beta}))$ if a logistic density was specified, and $\Phi(a_j - x_i\boldsymbol{\beta})$ if a Normal density was used where Φ is the cumulative distribution function of a standard Normal density. The logistic specification was favoured due to its simplicity in model parameter interpretation (Agresti, 2010). The probability of $(Y_i = j)$, conditioning on x_i, is $F(Y_i^* \le \alpha_j) - F(Y_i^* \le \alpha_{j-1})$. The GWOR models extended here allows for both Normal and logistic densities for ϵ.

The effect of a predictor variable, say x_1, on the cumulative probability of a response falling into category j is not linear because of the non-linear cumulative distribution function. This is seen from the partial derivative of the cumulative probability with respect to x_1, $\partial \mathrm{P}(Y_i \le j)/\partial x_1 = f(a_j - x_i\boldsymbol{\beta})\beta_1$ where $f(.) = F'(.)$ is the density function and β_1 the regression coefficient of x_1. The interpretation of estimated coefficients can make use of the concept of odds ratios as in a simple binary logit model. Taking the log odds of the cumulative probability in (2) and inserting the cumulative logistic probability formula, we obtain,

$$log\frac{\mathrm{P}(Y_i \le j)}{1 - \mathrm{P}(Y_i \le j)} = a_j - x_i\boldsymbol{\beta}. \tag{3}$$

The equation shows that the effect of x_1 on the cumulative probability on the logit scale is simply β_1 regardless the response category. The maximum likelihood estimation approach is usually used for model estimation. For observation i, let $y_{i1}, \dots, y_{i,J}$ be binary indicators of response categories, then we have $y_{ij} = 1$ and $y_{ik} = 0$ for $k \ne j$ if $Y_i = j$. The log-likelihood function of the model is,

$$l(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log[F(Y_i^* \le a_j) - F(Y_i^* \le a_{j-1})] \tag{4}$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \alpha_1, \dots, \alpha_{J-1}]$. It's useful to note that only $J - 1$ cut points are needed to divide the latent variable Y^* into J categories while α_0 and α_J are set to $-\infty$ and $+\infty$, respectively. Although there is not a tractable solution for the first-order conditions of the equation, it has been shown that the log-likelihood function has a unique global optimum so different types of iterative maximisation algorithms can be applied to estimate $\boldsymbol{\theta}$ (Burridge, 1981; Pratt, 1981).

3. A geographically weighted ordinal regression model

We now describe the geographically weighted ordinal regression model that allows for regression coefficients varying across space.