Geographically weighted modelling for spatial epidemiology

Tomoki Nakaya
Graduate School of Environmental Studies,
Tohoku University

1 Introduction

Spatial epidemiology is a research field seeking to obtain clues that aid disease control and health promotion through spatial data analysis (Elliott and Wartenberg, 2004). Geographic visualisation of epidemiological information, also known as disease mapping, plays an important role in this context. One example is Snow's pioneering study that identified a water pump as the plausible contamination source of a local cholera outbreak from the distribution of cases in the 19th century. Spatial epidemiology has been widely adopted and advanced with the advent of GIS (geographic information system) and related data analysis methods. In particular, local spatial analysis, which extracts statistical properties in geographically localised areas, has high affinities with GIS-based geographic visualisation. The developed methodology is also applicable to fields that focus on collective behaviours of events, such as crime mapping and analysis.

Popular tools include spatial scan statistics that explores areas with elevated risks of disease, the cartographic version of kernel density estimation (KDE) that estimates the smooth distribution of local disease density, and geographically weighted regression (GWR) that also uses spatial kernel to estimate smooth regional variations in relationships between variables (Fotheringham et al., 2002). Although GWR and KDE were not originally developed for epidemiological analysis, according to Pubmed, a literature database in the life science and health fields, the number of published articles containing these as keywords has increased rapidly since the 2000s, and more than 90 publications on GWR have been registered in 2019 (Figure 1). This short article is a summary of the author's talk in the symposium on 'Modelling spatial heterogeneity in environmental and ecological processes' which was held on 30 January 2020 at the Research Institute for Mathematical Sciences, Kyoto University. The aim of this article is to provide an overview of the basic concepts of GWR and its variants specifically with regards to spatial epidemiological studies, with a particular interest in how the methodology contributes to the creation of novel epidemiological mapping.

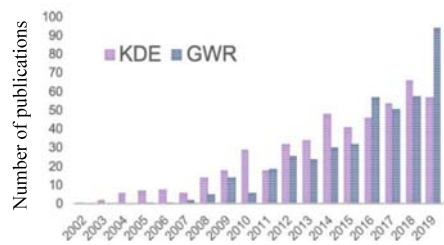


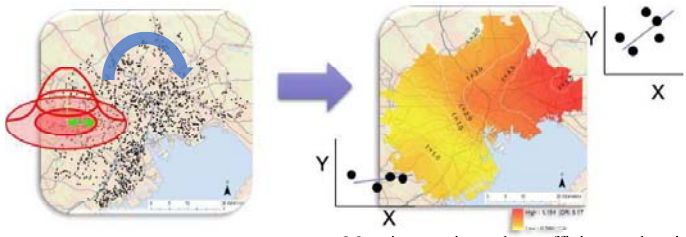
Figure 1 Number of publications with KDE or GWR registered in Pubmed

2 Concept of GWR

GWR is a type of conditional nonparametric regression that is formulated as a model with geographically varying coefficients:

$$y_i = \sum_k \beta_k(u_i, v_i) x_{k,i} + \varepsilon_i$$

where i is the subscript identifying the sample, y and x_k are the dependent variables and k-th explanatory



Repeating local fitting of regression model by moving a kernel.

Mapping estimated coefficients showing geographically varying X–Y relationships.

Figure 2 Concept of mapping geographically varying coefficient of GWR

variable and ε is the error term; $\beta_k(u_i, v_i)$ is a geographically local coefficient of the k-th variable and is assumed to fluctuate smoothly depending on the geographic location, (u_i, v_i). GWR estimates the local coefficients by repeatedly fitting the regression model to a weighted geographical subset of the data using a two-dimensional geographical kernel, giving more importance to locations that are nearer to the sample/regression point. The name of the method is derived from this geographical weighting approach of estimating geographically varying coefficients. Mapping estimated coefficients shows the variation in relationships between the dependent and explanatory variables (Figure 2).

GWR is the generalisation of geographically weighted generalised linear models (GWGLM) based on a geographically weighted maximum likelihood estimation (GW-MLE) which estimates the local coefficients by solving the following maximisation problem of the geographically weighted log-likelihood model for each point (Nakaya et al., 2005b; Nakaya, 2015):

$$\{\hat{\beta}_k(u_i, v_i)\} = \arg\max \sum_j \{log\, f\left(y_j | \eta_j(\{\hat{\beta}_k(u_i, v_i)\}), \hat{\varphi}\right) \cdot K(d_{ij}/h)\}$$

where the symbol ^ refers to the estimate and $\varphi$is the dispersion parameter; d_{ij} is the distance between locations i and j; the geographical weight of the j-th observation at the i-th regression point; K, is specified as a non-negative and monotonously decreasing function of the distance between the i-th and the j-th locations, like the Gaussian kernel function, $\exp(-0.5\, d_{ij}^2/h^2)$, where h is the bandwidth parameter controlling the local geographical extent of the weighting and the smoothness of estimated varying coefficients. Geographically weighted binary logistic and Poisson models are popular modes of GWGLM. Further, the idea of geographically weighted maximum likelihood estimation was extended to multi-response categorical models, such as geographically weighted ordered regression (Dong et al., 2018). Furthermore, it is possible to construct models mixing geographically varying and not-varying coefficients (semiparametric models) (Nakaya et al., 2015), or to introduce different degrees of smoothness (different bandwidth sizes) for each geographically varying term (multi-scale GWR) (Fotheringham et al., 2017).

3 Disease association mapping

The approach of GWR and its variants can be considered to be a natural extension of conventional KDE for disease mapping. While traditional disease mapping aims to reveal geographical variations of disease risks or health levels, GWR can be used for 'disease association mapping' to reveal geographical variations or spatial anomalies (clusters) in the relationship between health outcomes and various environmental variables. Historically, spatial epidemiologic analyses or studies of health geographies have often identified geographic contextuality that could alter the relationship between health outcomes and environmental variables

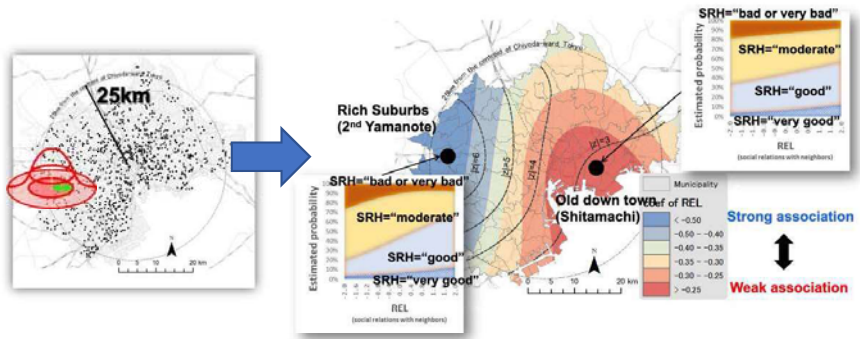


Figure 3 Disease association mapping using semiparametric geographically weighted ordered regression. The left-side map shows the sample residential locations and the right-side map shows the distribution of the estimated coefficients of neighbourhood social relationships that are used to predict individual self-rated health. The graphs inserted to the map show the postestimation of the relationships of women in their 40s with an average income.

reflecting different geographic backgrounds. For example, Kagami (1983) showed that cerebrovascular mortality tends to increase with lower winter temperatures in Japan, but not in Hokkaido, pointing out that the regional differences in living styles associated with indoor variations in temperature during the winter season could be a plausible reason for the geographically alternating relationship. In addition, the definition of the variables used as environmental indicators may vary geographically. For example, car ownership by households as an indirect indicator of household income which is often used in the UK may not be representative of income level in rural settings where automobile ownership is more of a necessity (Nakaya et al., 2005b). The random slope model, a type of multilevel model, was occasionally used for a similar purpose, estimating different coefficients between aggregated regions. While multilevel models require predefined geographic aggregation units for coefficients, GWR models do not require such prerequisite spatial conditions therefore they are more suitable for the exploratory data analysis of phenomena in which the relationship between health outcomes and environmental indicators varies according to the geographic context.

Figure 3 presents an example of the analysis of 1,089 adult women aged 30–59 years living in Tokyo special wards and their surroundings. Self-rated health (four category ordinal scales) is the outcome and a neighbourhood social relationship index is used as the explanatory variable.

Herein, the effects of income and age are adjusted by semiparametric models. The estimated result of the model indicated that the stronger the social relationship with the neighbourhood (the higher the value of the variable REL), the better the self-rated health of the individual. However, the relationship is weaker in areas close to the urban centre where redevelopment has progressed in recent years, and stronger in the western region which corresponds to socially established suburban residential areas. This relationship might be explained by the tendency for social relationships to be substituted by the growing number of residents expecting a highly anonymous lifestyle in highly urbanised areas, even though the relationship with neighbouring communities is weak.

4 GW modelling for space–time data analysis

This approach is applicable to spatiotemporal data analysis. For example, in the case of a multiregional susceptible-infective-recovered (SIR) model for the HIV epidemic in Japan, the regional parameters of

infectious contacts were estimated based on a non-linear variant of geographically weighted Poisson regression from the spatio-temporal series of the number of reported HIV cases in each prefecture (Nakaya et al., 2005a). The study revealed a distinctive regional variation in infectious contact rates of HIV.

Further, we can introduce geographically and temporally weighting regression (GWTR) to estimate spatio-temporal variations in coefficients of regression models (Huang et al., 2010). By combining GWGLM and GWTR, coefficients around spatiotemporal regression points (u_i,v_i,t_i) can be estimated based on the locally weighted maximum likelihood principle using geographical and temporal kernels simultaneously as follows:

$$\{\hat{\beta}_k(u_i, v_i, t_i)\} = \arg\max \sum_j \left\{ log\, f\left(y_j | \eta_j(\{\hat{\beta}_k(u_i, v_i)\}), \hat{\varphi}\right) \cdot K_s\left(\frac{u_j - u_i}{h_s}, \frac{v_j - v_i}{h_s}\right) K_t\left(\frac{t_j - t_i}{h_t}\right) \right\},$$

where K_s and K_t are kernel functions for the spatial and temporal domains, respectively, and h_s and h_t are their associated bandwidth parameters.

A special type of GWTR for spatial epidemiology that we proposed is GW-LOWESS (geographically weighted locally weighted scatterplot smoothing), which involves extending the concept of space–time KDE or GWTR to capture localised anomalous emerging trends of disease incidence (Nakaya et al. 2014). Considering the case that disease occurrences are recorded in spatio-temporal aggregated units, a variant of GW-LOWESS based on a Poisson regression scheme was proposed as:

$$y_i \sim Poisson[\mu_i]$$
$$\mu_i = A_i\, exp(\beta_0(u_i, v_i, t_i) + \beta_t(u_i, v_i, t_i) t_i),$$

where μ_i is the expected count of events and t_i is the continuous variable of time; A_i is the offset variable of observation i used for adjusting the size of the observation unit, such as the areal size or population at risk. A simple way to assess localised emerging trends is attained by using local testing of the temporal coefficient, $\beta_t(u_i, v_i, t_i)$, with the Wald statistic.

Nakaya et al. (2014) applied this technique to a dataset of 611 reported crimes as a social disease (snatch-and-run crime) in the central part of Osaka City, Japan. The dataset had a monthly temporal resolution covering a two-year period, 2011–2012. Figure 4 shows the spatio-temporal distribution of reported crimes in a space–time cube on the left-side and the space–time domain having high positive Wald statistics (z = 1.96 and z = 2.50) of temporal coefficients of GW-LOWESS on the right-side. Highly positive Wald statistics

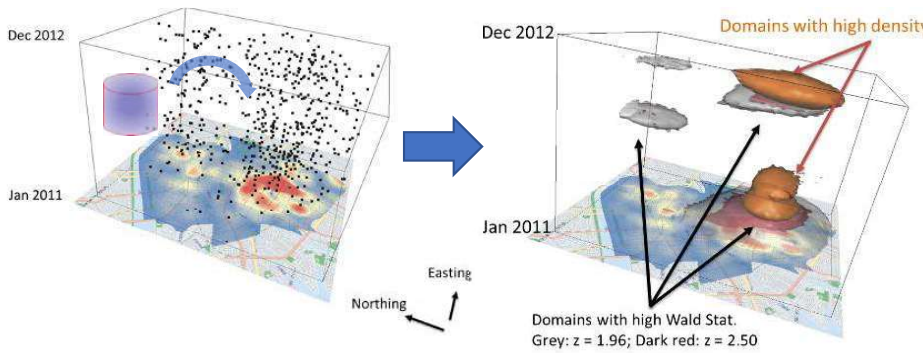


Figure 4 GW-LOWESS for detecting emerging clusters
The left-side image shows the space–time distribution of crime incidence in a space–time domain. A space–time kernel is represented as a cylinder like shape. The right-side image shows space-time contours of domains with emerging trends (coloured in grey) with domains with high density (coloured in orange).

were notable before the space–time domains with high crime density emerged in the south part of the region.

This result indicated that GW-LOWESS successfully captured the emerging trends of incidence and can be used for the early detection of emerging crime/disease clusters.

5 Conclusions

Geographically weighted modelling makes a unique contribution to spatial epidemiology through various styles of association mapping of diseases as novel modes of health geo-visualisation. Standard GWR models can be used by R packages like GWmodel, dedicated software (GWR4) (Nakaya, 2015), and inbuilt functions in commercial GIS's, like ArcGIS (ESRI Inc.). In addition, this approach is theoretically simple to modify for generating new forms of models and visualisation methods that are required to meet the challenges of various epidemiological analyses. A few examples of such efforts were demonstrated in this article. Further efforts are expected to be made in terms of robust estimation and fast computing of geographically large-scale data in epidemiology.

References

Dong, G., Nakaya, T. and Brunsdon, C. (2018): Geographically weighted regression models for ordinal categorical response variables: An application to geo-referenced life satisfaction data. *Computers, Environment and Urban Systems*, 70, 35-42.

Elliott, P. and Wartenberg, D. (2004): Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, 112, 998-1006.

Fotheringham, A. S., Yang, W., and Kang, W. (2017): Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107, 1247-1265.

Kagami, M. (1983): Regional variance of cerebrovascular mortality in Japan. *Ecology of Disease*, 2, 277-283.

Huang, B., Wu, B. and Barry, M. (2010): Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24, 383-401.

Nakaya, T., Nakase, K and Osaka, K. (2005a): Spatio-temporal modelling of the HIV Epidemic in Japan based on the national HIV/AIDS surveillance. *Journal of Geographical Systems*, 7, 313-336.

Nakaya, T., Fotheringham, S., Brunsdon, C. and Charlton, M. (2005b): Geographically weighted Poisson regression for disease association mapping, *Statistics in Medicine* 24, 2695-2717.

Nakaya T, Haworth J, Cheng T (2014): Visualising Emerging Trends of Clusters in a Space-Time Region Using Spatio-Temporal Kernel Regression. In Stewart K, Pebesma E, Navratil G, Fogliaroni P, Duckham M eds. *Proceedings of GIScience 2014*, 200-204.

Nakaya, T. (2015): Geographically weighted generalised linear modeling. Brunsdon, C. and Singleton, A. eds. *Geocomputation: A Practical Primer*', Sage Publication, 201-220.

Graduate School of Environmental Science, Tohoku University
Sendai, Miyagi, 980-0845, Japan
Email to: tomoki.nakaya.c8@tohoku.ac.jp

This work was supported by JSPS KAKENHI 16H01830 and the Joint Support Center for Data Science Research (ROIS-DS-JOINT) under Grant 004RP2019.