Spatial Statistics 23 (2018) 36–51

Contents lists available at ScienceDirect

Spatial Statistics

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

A hierarchically adaptable spatial regression model to link aggregated health data and environmental data

Phuong N. Truong [*,1], Alfred Stein

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 16 January 2017
Accepted 6 November 2017
Available online 15 November 2017

Keywords:
Health geography
Aggregated health data
CoS
Ecological fallacy
Spatially varying coefficient
HFMD

ABSTRACT

Health data and environmental data are commonly collected at different levels of aggregation. A persistent challenge of using a spatial regression model to link these data is that their associations can vary as a function of aggregation. This results into ecological fallacy if association at one aggregation level is used for inferencing at another level. We address this challenge by presenting a hierarchically adaptable spatial regression model. In essence, the model extends the spatially varying coefficient model to allow the response to be count data at larger aggregation levels than that of the covariates. A Bayesian hierarchical approach is used for inferencing the model parameters. Robust inference and optimal prediction over geographical space and at different spatial aggregation levels are studied by simulated data sets. The spatial associations at different spatial supports are largely different, but can be efficiently inferred when prior knowledge of the associations is available. The model is applied to study hand, foot and mouth disease (HFMD) in Da Nang city, Viet Nam. Decrease in vegetated areas corresponds with elevated HFMD risks. A study to the identifiability of the parameters shows a strong need for a highly informative prior distribution. We conclude that the model is robust to the underlying aggregation levels of the calibrating data for association inference and it is ready for application in health geography.

* Corresponding author.
E-mail address: p.truongngocphuong@utwente.nl (P.N. Truong).
[1] Postal address: PO Box 217, 7500 AE Enschede, The Netherlands.
Abbreviations: SLM: spatial multilevel statistical model

https://doi.org/10.1016/j.spasta.2017.11.002
2211-6753/

1. Introduction

Important scientific research questions in health geography to study the effects of environmental exposure on human health are "What is the association between human health and the environment?" and "Which of the associations is statistically significant?" (Steenland and Dedden, 1997; Pekkanen and Pearce, 2001). In many cases, for the reasons of confidentiality and cost, extensive geo-referenced health data are only accessible as aggregated data at administrative levels. Meanwhile, environmental data are collected from stations monitoring air, soil or water at a relatively low number of locations or over smaller areas in order to capture their usually small scale variations. Remote sensing imagery such as MODIS, LANDSAT, etc. provides geographically extensive data of the environment but at various spatial resolutions. Unsurprisingly, in many circumstances, these data do not immediately have comparable spatial measurement units in terms of resolution or spatial support to health data. For example, provincial medical statistics in Viet Nam only report hand, food and mouth disease (HFMD) cases at the district level with area sizes ranging between 10 and 10^3 km^2; whereas the environmental risk factors for this disease such as daily air temperature and humidity are regularly recorded at only one or two meteorological monitoring stations per province with an average area of about 5×10^3 km^2. Persistent challenges to link these data are that their associations can vary as a function of aggregation, the well-known modifiable areal unit problem (MAUP) (Openshaw, 1983; Cressie, 1996). This results in ecological fallacy if association at one aggregation level is carelessly used for inferencing at another aggregation level (King, 1997).

Popular models for association analyses in health geography are the regression-based models (Keppel, 2005; Bender, 2009; Auchincloss et al., 2012). A well-known example of such models in health research is the spatial multilevel statistical model (SLM) (Langford et al., 1999; Goldstein, 2010; Arcaya et al., 2012). Here, the "spatial" prefix distinguishes the hierarchy of the geographical space from the hierarchy of the feature space of the data. Spatial hierarchy is defined by the differences of the spatial supports. For example, morbidity reporting of individual person contracted HFMD in Viet Nam (level 1) is aggregated to district level (level 2) and to regional level (level 3). SLM is widely applied to health data with the spatial hierarchy structuring from individual health outcomes to larger environmental surroundings. Such a desired spatial hierarchy, however, does not always hold in practice due to lack of exclusive sampling designs (Duncan and Jones, 2000).

The associations between health data and environmental risk factors can locally vary due to various facts such as averaging effects of aggregated health data, measurement units of environmental risk factors, spatially changing socio-economic and individual characteristics of the understudied population. Also, the spatial autocorrelation of health data, environmental risk factors or both have an effect. Previous research (e.g. Fotheringham et al., 1998, 2001; Saib et al., 2014; Hamm et al., 2015, amongst others) has demonstrated the superiority of geographically adaptable regression coefficient models in various real cases if spatial non-stationarity of the associations is present. The two best known models in current literature are the geographically weighted regression model (GWR) (Fotheringham et al., 1998, 2009) and the spatial varying coefficient model (SVC) (Gelfand et al., 2003). In GWR, the parameter surface is allowed to vary as a deterministic spatial surface; and its value at an individual location is estimated by geographical distance weighted least square of proximate locations (Fotheringham et al., 2001). In SVC, it is generated by a second-order stationary spatial random process that enables the probabilistic uncertainty quantification about the parameter estimators (Gelfand et al., 2003).

The underlying principle of both GWR and SVC is that the regression coefficient surfaces follow Tobler's first law of geography, i.e. that the regression coefficients are spatially correlated. GWR, however, is not robust to the MAUP (Fotheringham et al., 2001). Change of support (CoS) modelling (Cressie, 2015a) is important in spatial statistics and SVC has been shown to be efficient in modelling spatial non-stationarity of the regression coefficients (Wheeler and Calder, 2007; Finley, 2011). Therefore, we argue that SVC provides an efficient modelling framework to investigate the effect of CoS on the parameters of the regression-based model for geo-referenced data with different spatial supports. Nevertheless, the increasing number of the parameters of the SVC might pose other challenges, e.g. to inference the hyper-parameters of the unobservable stochastic regression coefficient surfaces. This identifiability issue has been experienced by many complex (spatial-temporal) statistical models (Brun et al., 2001; Bujosa et al., 2007; Lavielle and Aarons, 2016; Ugarte et al., 2017).