Balancing spatial and non-spatial heterogeneity in large samples

Daisuke Murakami
Department of Statistical Data Science,
Institute of Statistical Mathematics

1 Introduction

An increasing number of geo-spatial data are becoming available in these days. Such data includes climate-related data and land cover data observed from satellites, socio-economic data by municipal units, and people flow data observed through human sensors. Regression is widely applied to these spatial data to reveal patterns behind geographical phenomena such as disease spread, economic development, and extinction of life in each region.

In spatial regression, spatial heterogeneity is considered as a key factor (Anselin, 2010). Spatial heterogeneity means that parameters characterizing spatial phenomena can vary over space (see, Brunsdon et al., 1998). For example, Figure 1 shows correlation plots between median household income and the ratio of residents speaking English (Eng_rat). This figure shows that Eng_rat has a strong positive impact on income in the California state, a moderate positive impact in the New York state, and a weak negative impact in the Louisiana state. In other words, relationship between variables vary over space. To capture such spatial heterogeneity, we need to allow for regression coefficients to vary over space.

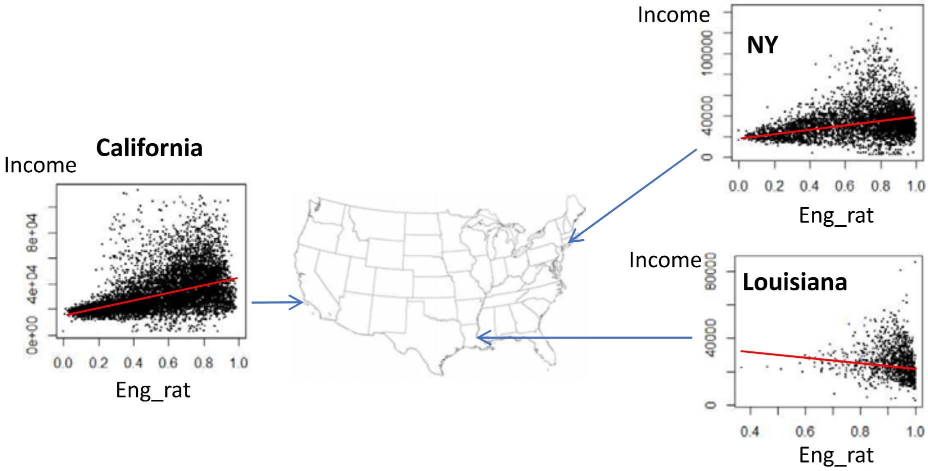

Figure 1 Correlation plots between Eng_rat and income in three states.

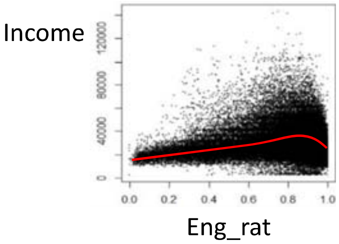

Figure 2 Correlation plots between Eng_rat and income by counties across the US.

Regression coefficients can also vary non-spatially. For example, Figure 2 is the correlation plot between Eng_rat and income over the United States (US). As illustrated in this figure, there is a positive linear relationship between these two variables if Eng_rat is between 0.0 and 0.9 whereas they have weak negative relationship if Eng_rat is between 0.9 and 1.0. The importance of considering such such non-spatial heterogeneity depending on covariate value has been suggested by Hastie and Tibshirani (1993) and many other studies.

In short, consideration of spatial and non-spatial heterogeneity is required to appropriately analyze geographical phenomena. Given that, this study develops a regression approach estimating spatially varying coefficients (SVCs) capturing spatial heterogeneity, and non-spatially varying coefficients (NVCs) capturing non-spatial heterogeneity. Because spatial data are getting bigger and bigger recently, we reduce the computational cost as much as possible to make it work for large samples.

2 Spatially and non-spatially varying coefficient (S&NVC) model

We model the explained variable y_i observed at the i-th sample site using the following spatially and non-spatially varying coefficient (SNVC) model (see Murakami and Griffith, 2020):

$$y_i = \sum_{k=1}^{K} x_{i,k}\beta_{i,k} + \varepsilon_i, \qquad \beta_{i,k} = b_k + f_{s,k}(s_i;\boldsymbol{\theta}_i) + f_n\left(x_{i,k};\boldsymbol{\varphi}_i\right), \qquad \varepsilon_i \sim N(0,\sigma^2), \tag{1}$$

where $x_{i,k}$ is the k-th covariate, and ε_i is a disturbance with variance σ^2. This model defines the k-th regression coefficient $\beta_{i,k}$ at i-th site s_i by [constant: b_k] + [SVC: $f_{s,k}(s_i;\boldsymbol{\theta}_i)$] + [NVC: $f_n(x_{i,k};\boldsymbol{\varphi}_i)$]. The k-th SVC is defined by a spatial process $f_{s,k}(s_i;\boldsymbol{\theta}_i)$, which has a smooth map pattern. Likewise, the k-th NVC is defined by a function $f_n(x_{i,k};\boldsymbol{\varphi}_i)$, whose value varies smoothly depending on $x_{i,k}$; the natural spline generated from $x_{i,k}$ is employed to the function. Each SVC process is defined by a linear combinations L ($< N$) spatial basis functions whereas each NVC is defined by the same with L_k ($< N$) natural spline functions. $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_K\}$ are sets of variance parameters specifying structures of SVC and NVC. This model attempts to balance constant, SVC, and NVC in each of the regression coefficients by estimating these variance parameters.

Unfortunately, estimation of the variance parameters is slow if for large samples. Roughly speaking, this is because we need to iteratively process a large matrix with dimension being $N \times (LK + \sum_{k=1}^{K} L_k)$ when estimating the variance parameters where N is the sample size. To lighten the cost, this study develops a fast restricted maximum likelihood (REML) estimation approach estimating the parameters by first replacing the large matrix with a $(LK + \sum_{k=1}^{K} L_k) \times (LK + \sum_{k=1}^{K} L_k)$ matrix of inner products, and iteratively process the small matrix instead of the large matrix; because dimension of the inner product matrix is independent of the sample size (N), the estimation is done computationally quite efficiently even for large samples.

In summary, the S&NVC model estimates or balances SVC and NVC computationally efficiently.

3 Application

We applied the developed S&NVC model to the Lucas housing price dataset with sample size of 25,357. The explained variables are lagged housing prices and the covariates are building age (Age), number of rooms (Rooms), and number of beds (Beds).

The model estimation took 142.655 seconds. Table 1 summarizes estimated variance of the spatial and non-spatial variations in each regression coefficient. This table suggest that influence from Age varying spatially and non-spatially, the influence from Room varying spatially, and the influence from Bets varying non-spatially.

Figure 3 plots the estimated coefficients (SVC + NVC). The estimated intercept suggested lower house price in the center. The estimated coefficients on Age demonstrated that oldness of residence is a strong negative factor in the city center located in the center of the map. While Rooms have positive impact on prices across the region, the influence is strong especially in the suburbs. Finally, influence from Beds have positive impact, and the impact gets strong in the center and the eastern area.

Table 1 Estimated variance of the spatial (SVC) and non-spatial variations (NVC) in each $\beta_{i,k}$. We assume spatial variation in the intercept.

	Intercept	Age	Rooms	Beds
SVC	0.048	0.073	0.006	0.000
NVC	N.A.	0.099	0.000	0.007

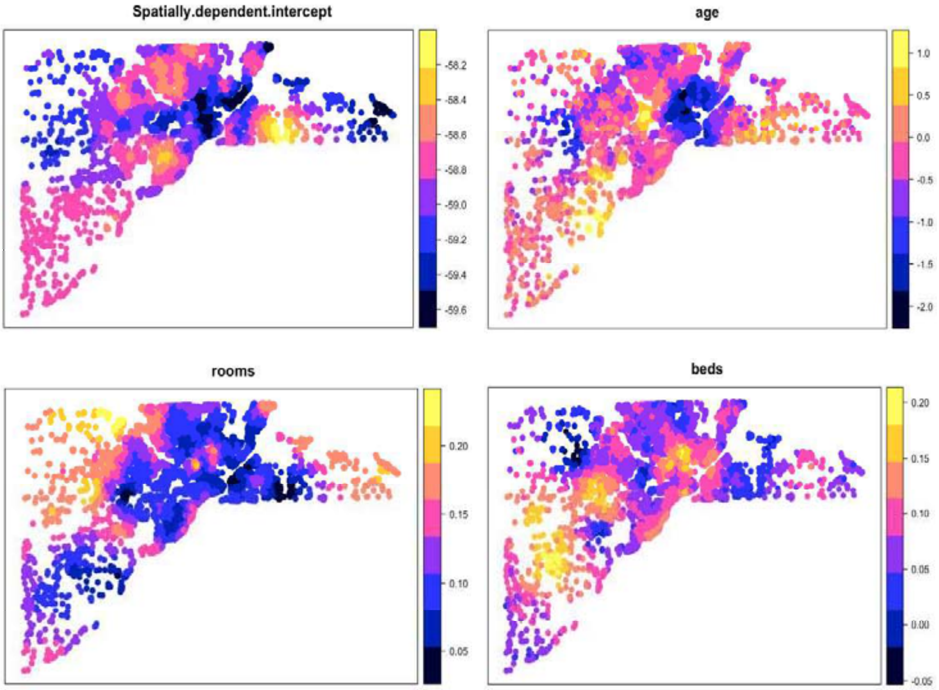

Figure 3 Estimated coefficients

4 Concluding remarks

This study explains the importance of considering spatial variation and non-spatial variation in regression coefficients, and empirically demonstrated that these two variations are present in real data. Approaches considering such variation includes geographically weighted regression models (Brunsdon et al., 1998), latent Gaussian models (e.g., Rue et al., 2009), and our spatial additive mixed models. These approaches will be further important together with the increase of available spatial dataset.

The S&NVC model is implemented in an R package spmoran (Murakami, 2020; https://cran.r-project.org/web/packages/spmoran/index.html).

Reference

Anselin, L. (2010): Thirty years of spatial econometrics. *Papers in regional science*, 89, 3-25.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998): Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 431-443.

Hastie, T., and Tibshirani, R. (1993): Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55, 757-779.

Murakami, D. (2017): spmoran (ver. 0.2.0): An R package for Moran eigenvector-based scalable spatial additive mixed modeling. *ArXiv*, 1703.04467.

Murakami, D. and Griffith, D. A. (2020): Balancing spatial and non-spatial variation in varying coefficient modeling: a remedy for spurious correlation. *ArXiv*, 2005.09981.

Rue, H., Martino, S., and Chopin, N. (2009): Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71, 319-392.

Department of Statistical Data Science, Institute of Statistical Mathematics
Midori-cho, Tachikawa, 190-8562, Japan
Email to: dmuraka@ism.ac.jp

This work was supported by JSPS KAKENHI 20K13261 and the Joint Support Center for Data Science Research (ROIS-DS-JOINT) under Grant 004RP2019.