Title

Geographically weighted regression for compositional data: An application to the U.S. household income compositions

Authors

Yoshida, Takahiro
Murakami, Daisuke
Seya, Hajime
et al.

Publication Date

2021-09-01

DOI

10.25436/E2G599

Peer reviewed

Geographically weighted regression for compositional data: An application to the U.S. household income compositions

Takahiro Yoshida[1]
The University of Tokyo, Japan
yoshida.takahiro@up.t.u-tokyo.ac.jp

Daisuke Murakami
Institute of Statistical Mathematics, Japan
dmuraka@ism.ac.jp

Hajime Seya
Kobe University, Japan
hseya@people.kobe-u.ac.jp

Narumasa Tsutsumida
Saitama University, Japan
narut@mail.saitama-u.ac.jp

Tomoki Nakaya
Tohoku University, Japan
tomoki.nakaya.c8@tohoku.ac.jp

Abstract

This study builds a bridge between the literatures for geographically weighted regression (GWR) and compositional data analysis (CoDA). GWR allows the modeling of spatial heterogeneity in regression models and is increasingly used in various fields. CoDA provides unique and useful tools for compositional data, which are restricted by a constant-sum constraint. Although compositional data are common in many scientific areas, it is not until recently that increasingly sophisticated statistical methods have been deeply investigated. Many types of spatial models based on geostatistics, spatial statistics, and spatial econometrics for compositional data have been proposed. However, there is less attention to both spatial heterogeneity and the constant-sum constraint. In this study, we propose a GWR model for compositional data. This allows us to model spatially varying relationships while considering the constant-sum constraint. We applied this model to analyze household income compositions at the county level in the US. The interpretational usefulness of the results of spatially varying compositional semi-elasticities is empirically performed.

Funding This research was supported by the ROIS-DS-JOINT Grant number 003RP2020 and the JSPS KAKENHI Grant Numbers 18H03628, 21H01447, and 21K13153.

---

[1] corresponding author

1 Introduction

Geographically weighted regression (GWR) [2] has been widely used in various fields. The extension for non-Gaussian distributed data has also been progressing. However, studies on the extension for compositional data, which are restricted by a constant-sum constraint such as 1 for proportions and 100 for percentages, are quite limited.

Although compositional data are commonly found in various scientific areas, it has not been until recently that the statistical analysis for compositional data, typically termed compositional data analysis (CoDA) [1, 5], has gained momentum. Currently, the development of spatial regression models for compositional data is one of hot topics in the CoDA literature. Geostatistical compositional models such as compositional kriging is popular approaches because CoDA are historically developed in geosciences in which a continuous spatial process can naturally be assumed. In other words, models with a discrete spatial process are relatively limited. Some papers employ conditional autoregressive models [9] or simultaneous autoregressive (spatial econometric) models [8]. In these models, spatial auto-correlation are considered. However, models for compositional data with spatial heterogeneity or spatially varying relationships are still quite limited.

The objective of this study is to propose a GWR model for compositional data to consider spatial heterogeneity and the constant-sum constraint. We accommodate the GWR model and logratio techniques of CoDA, and then formulate the GWR model in the simplex space, which is the sample space of compositional data.

2 Fundamental concepts and operators of CoDA

2.1 Aitchison geometry in the simplex space

A vector $\mathbf{p} = (p_1, p_2, \dots, p_D)$ whose components are positive real numbers and carry relative information is called as a D-part composition. The composition can be represented as an element of the D-part simplex space $\mathcal{S}^D$:

$$\mathcal{S}^D = \left\{ \mathbf{p} = (p_1, p_2, \dots, p_D) \;\middle|\; p_m > 0, m = 1, 2, \dots, D, \sum_{m=1}^{D} p_m = \kappa \right\}, \tag{1}$$

where κ is a constant sum for compositions in $\mathcal{S}^D$. Usual values of κ are 1 (proportions) and 100 (percentages: %). Rescaling of compositions can be formalized by the closure operator $\mathcal{C}_\kappa$ for $\mathbf{z} = (z_1, z_2, \dots, z_D) \in \mathbb{R}_+^D$: $\mathcal{C}_\kappa(\mathbf{z}) = \left(\frac{\kappa \cdot z_1}{\sum_{m=1}^{D} z_m}, \frac{\kappa \cdot z_2}{\sum_{m=1}^{D} z_m}, \dots, \frac{\kappa \cdot z_D}{\sum_{m=1}^{D} z_m} \right)$. The constant-sum constraint induces statistical problems such as the restriction of the degree of freedom and the spurious correlation for the use of standard statistical methods with compositions [1].

The geometrical structure of compositions has been established to define a vector space structure of the simplex space, and it is reffered as the Aitchison geometry. The two basic operations are the pertubation operator and the powering operator which correspond to the addition/shifting operator and the multiplication operator in the Euclidean geometry, respectively. For two D-part compositions $\mathbf{p}, \mathbf{q} \in \mathcal{S}^D$ and a constant scalar $\alpha \in \mathbb{R}$, the pertubation operator $\oplus$ is: $\mathbf{p} \oplus \mathbf{q} = \mathcal{C}_\kappa(p_1 \cdot q_1, p_2 \cdot q_2, \dots, p_D \cdot q_D) \in \mathcal{S}^D$ and the power operator $\odot$ is: $\alpha \odot \mathbf{p} = \mathcal{C}_\kappa\left(p_1^\alpha, p_2^\alpha, \dots, p_D^\alpha\right) \in \mathcal{S}^D$. By using the two fundamental operators, for $\mathbf{p}_k \in \mathcal{S}^D$, $\alpha_k \in \mathbb{R}$, $k = 1, 2, \dots, K$, the purtubation-linear combination operator $\bigoplus$ is introduced: $\bigoplus_{k=1}^{K} \alpha_k \odot \mathbf{p}_k = \alpha_1 \odot \mathbf{p}_1 \oplus \alpha_2 \odot \mathbf{p}_2 \oplus \cdots \oplus \alpha_K \odot \mathbf{p}_K = \mathcal{C}_\kappa\left(\prod_{k=1}^{K} p_{k,1}^{\alpha_k}, \prod_{k=1}^{K} p_{k,2}^{\alpha_k}, \dots, \prod_{k=1}^{K} p_{k,D}^{\alpha_k} \right) \in \mathcal{S}^D$.

2.2 Logratio transformation

Since most standard statistical methods depend on the Euclidean geometry in the real space, it is reasonable to project compositions from the simplex to the real space. To construct such projections, some transformations have been proposed. Classical transformations are the additive logratio (alr) [1]; the centered logratio (clr) [1]; and the isometric logratio (ilr) [6]. It can be said that the CoDA literature has been discussing and providing the general framework of the logratio transformation. In this paper, the ilr transformation is used because it is based on an orthonormal basis, so that it is well recognized as the most preferable from a mathematical point of view. There are infinitely many possibilities to define such an orthnormal basis. In the CoDA literature, the following ilr orthnormal coordinates referred to as the pivot coordinates [7] is currently used as a preferable option. The ilr transformation with the pivot coordinates for $\mathbf{p} \in \mathcal{S}^D$ is defined as follows: $\mathrm{ilr}(\mathbf{p}) = \mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_{D-1}^*) \in \mathbb{R}^{D-1}$ with $p_l^* = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{p_j}{\sqrt[D-j]{\prod_{\ddot{m}=1}^{D} p_{\ddot{m}}}}, l = 1, 2, \ldots, (D-1)$, where superscript $*$ denotes the ilr transformation.

3 GWR model for compositional data

The GWR model is an extention of the linear regression model that allows regression coefficients to vary across geographical space. When the explained variable is a D-part composition, the basic GWR model in the simplex $\mathcal{S}^D$ can be expressed as follows:

$$\mathbf{y}_i = \bigoplus_{k=1}^{K+1} (x_{i,k} \odot (\boldsymbol{\beta}_i)_k) \oplus \boldsymbol{\varepsilon}_i, \tag{2}$$

where $i \in \{1, 2, \ldots, n\}$ is the index for sites; $\mathbf{y}_i$ is the explained variables; $x_{i,k}$ is the k-th covariate; $K+1$ is the number of covariates including intercept; $(\boldsymbol{\beta}_i)_k$ is unknown parameters of $x_{i,k}$; $\boldsymbol{\varepsilon}_i$ is the disturbances. $\mathbf{y}_i$, $(\boldsymbol{\beta}_i)_k$, and $\boldsymbol{\varepsilon}_i$ are D-part compositions in $\mathcal{S}^D$. In order to estimate parameters of the model, we consider the following two characteristics: (1) constant-sum constraint and (2) spatial heterogeneity.

STEP 1: For considering the constant-sum constraint of compositional explained variables and obtaining the model in real space, we use the ilr transformation. The ilr transformed model for the i-th observation site in the l-th GWR model as a scalar representation can be expressed as $y_i^{*(l)} = \sum_{k=1}^{K+1} \left(x_{i,k} \cdot \left(\beta_i^{*(l)} \right)_k \right) + \varepsilon_i^{*(l)}$.

STEP 2: Each transformed model can be estimated independently [4]. Therefore, the estimation of the basic GWR model can be applied. Thus, the regression coefficients $\boldsymbol{\beta}_i^{*(l)}$ is given by the weighted least squares estimators as: $\hat{\boldsymbol{\beta}}_i^{*(l)} = \left[\mathbf{X}'\mathbf{G}_i(b^{(l)})\mathbf{X} \right]^{-1} \mathbf{X}'\mathbf{G}_i(b^{(l)})\mathbf{y}^{*(l)}$, where $\mathbf{X}$ is the covarites matrix whose (i,k)-th element equals $x_{i,k}$, $\mathbf{G}_i(b^{(l)})$ is an $n \times n$ diagonal matrix, whose j-th element assigns the weight on the j-th sample site. The weight is given by a distance-decaying kernel, which we assumed the Gaussian kernel. $b^{(l)}$ is the kernel bandwidth, which can vary for each l. The GWR model can be estimated by first optimizing the bandwidth, and estimating the regression coefficients $\boldsymbol{\beta}_i^{*[l]}$ after that. The bandwidth can be optimized by the leave-one-out cross-validation method to minimize the cross-validation score.

4 Empirical analysis

4.1 Outline

This section illustrates an application of out proposed model to the United States (US) household income dataset 2017. The explained variable is county-level compositional household income data divided into the high-income bracket with households whose income in the past 12 months was more than \$75,000, middle-income bracket with households earning between \$35,000–\$75,000, and low-income bracket of less than \$35,000. The sample size is 3,108. To maintain a continuous geographical space, Alaska and Hawaii are excluded from the sample. To discuss regional differences, we selected three counties: New York County, New York (NY), El Paso County, Texas (TX), and Park County, Wyoming (WY). The covariates are as follows: *Univ* is the percentage of people with a bachelor's degree or higher among people over 25 years old, *Eng* is the percentage of people who speak English, and *Age* is the median age.

4.2 Results

Figure 1 summarizes the estimated semi-elasticities of each covariate for each bracket. Because the dependent variable is transformed, it is not appropriate to directly interpret and visualize the regression coefficients. The semi-elasticity gives the relative percentage change in the dependent variable when the covariate increases by 1 unit. It is noted that the sum of the semi-elasticities for each bracket in the compositional model is 0. This helps us easily and directly interpret the impact. Additionally, thanks to the GWR model, the semi-elasticity spatially varies. For example, when the covariate *Univ* of New York County increases by 1 unit, high income changes +0.513%, middle income –0.256%, and low income –0.257%. In the same way, for El Paso County, high income changes +0.509%, middle income –0.006%, and low income –0.443%. For Park County, high income changes +0.282%, middle income –0.151%, and low income –0.131%. As a result, the impact on the low-income households of New York County is about two times that of Park County. From the spatial distributions in Figure 1, *Univ* has a positive impact on the high-income bracket, especially in metropolitan counties on the east and west coasts. Because there are many white-collar and professional workers living in these counties, this result is reasonable. *Age* has a positive impact on the low-income bracket. *Age* also has a positive impact on the high-income bracket of some counties in the eastern area. Based on the results, older veteran workers have higher earnings in these counties. *Eng* has a strong impact on each bracket. In the northwestern area, *Eng* has a strong positive impact on the high-income bracket and a strong negative impact on the low-income bracket. In the southern area, which is close to the Mexico-US border, speaking English appears to have a positive impact on the high-income bracket.

Figure 2 illustrates the effects of the inverse-transformed estimated coefficients, in which the change in the predicted probabilities for each bracket can be seen as a function of the change in the covariate level. When we hold the non-target covariates at the observed values, we can examine the predicted probabilities across the observed range of each covariate individually. In the model, the predicted probabilities are also spatially varying, so the results can be comparable among sites. From the comparison of the three counties, *Univ* shows a positive impact regarding high-income households, with the strongest relationship occurring in New York County. When *Univ* is around 8% – 10%, the most dominant bracket changes. Among the three counties, *Age* and *Eng* exhibit different patterns. In New York County, *Age* does not affect much change. *Age* has a linearly increasing effect on the low-income bracket

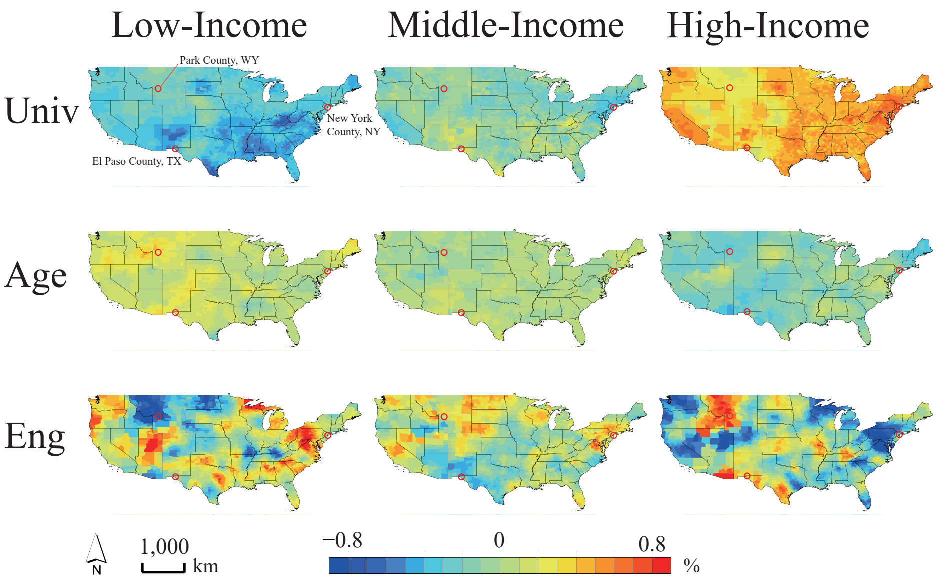

Figure 1 Semi-elasticity of each covariate for each bracket.

in El Paso and an exponential effect in Park. Although *Eng* in New York does not change the income brackets much, the high-income brackets increase in El Paso and Park. In Park, *Eng* exponentially decreases the low-income bracket.

In summary, the study provides an empirical evidence that our proposed model successfully captures spatial patterns in the regression results. Although the regression coefficients cannot be interpreted, the semi-elasticities and predicted probabilities are directly and easily interpretable. The model can be useful for a wide variety of spatial modeling with spatial heterogeneity and compositional characteristics.

5 Discussion and conclusion

This study aims to develop a methodology for geographically weighted regression (GWR) for compositional data that models spatially varying coefficients restricted in a simplex space. These findings are meaningful because spatial compositional data are common in many fields, including environmental sciences and geography. An analysis of household income compositional data in the United States demonstrated that spatially varying compositional semi-elasticities with a sum restricted to 0 and spatially varying predicted probabilities provide insights into a spatial non-stationary phenomenon.

Our proposed model can be considered in the extension of GWR modeling for non-Gaussian distributed data, which has been progressing in the spatial analysis literature. [3] proposes a geographically weighted beta regression for a rate or proportion that is usually defined between (0, 1). Naturally, one potential extension is to consider Dirichlet distributed data. Developing a geographically weighted Dirichlet regression and comparing it with our approach is an interesting topic for future research.

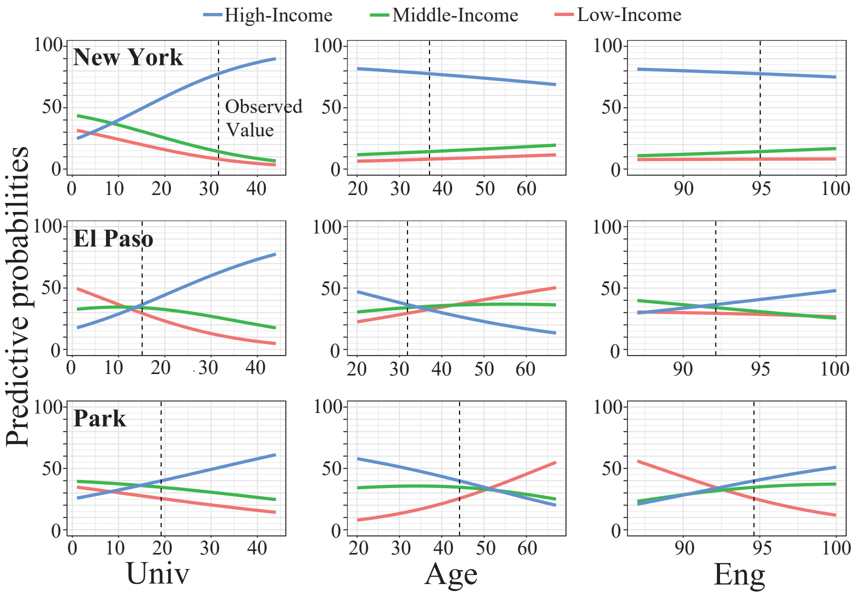

Figure 2 Predictive probabilities for each bracket regarding each covariate for New York County, NY (top), El Paso County, TX (middle), and Park County, WY (bottom). In each panel, the target covariate varies across the observed range of data and the non-target covariates are held at the observed values.

References

1 John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

2 Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996.

3 Alan Ricardo da Silva and Andreza de Oliveira Lima. Geographically weighted beta regression. *Spatial Statistics*, 21:279–303, 2017.

4 Juan José Egozcue, Carles Barceló-Vidal, Josep Antoni Martín-Ferńandez, Eusebi Jarauta-Bragulat, José Luis Díaz-Barrero, and Glòria Mateu-Figueras. Elements of Simplicial Linear Algebra and Geometry. *In: Compositional Data Analysis: Theory and Applications (Eds: Vera Pawlowsky-Glahn and Antonella Buccianti)*, pages 139–157, 2011.

5 Juan José Egozcue and Vera Pawlowsky-Glahn. Compositional data: the sample space and its structure. *Test*, 28(3):599–638, 2019.

6 Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

7 Eva Fišerová and Karel Hron. On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43(4):455–468, 2011.

8 Thi Huong An Nguyen, Christine Thomas-Agnan, Thibault Laurent, and Anne Ruiz-Gazen. A simultaneous spatial autoregressive model for compositional data. *Spatial Economic Analysis*, 16(2):161–175, 2021.

9 Takahiro Yoshida and Morito Tsutsumi. On the effects of spatial relationships in spatial compositional multivariate models. *Letters in Spatial and Resource Sciences*, 11(1):57–70, 2018.