


Moran Eigenvectors-Based Spatial Heterogeneity Analysis for Compositional Data

Zhan Peng ✉ 

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Ryo Inoue ✉ 

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Abstract

Spatial analysis of data with compositional structure has gained increasing attention in recent years. However, the spatial heterogeneity of compositional data has not been widely discussed. This study developed a Moran eigenvectors-based spatial heterogeneity analysis framework to investigate the spatially varying relationships between the compositional dependent variable and real-value covariates. The proposed method was applied to municipal-level household income data in Tokyo, Japan in 2018.

2012 ACM Subject Classification Applied computing → Mathematics and statistics

Keywords and phrases Compositional data analysis, Spatial heterogeneity, Moran eigenvectors

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.59

Category Short Paper

Funding This study was supported by JSPS KAKENHI Grant Number JP21H01447 and JST SPRING Grant Number JPMJSP2114.

1 Introduction

Spatial data that represent parts of a whole and carry only relative information are known as compositional data, such as income structure, land use shares, and vote shares across multiple regions. Although previous studies have considered both the compositional and spatial nature of data [5], little attention has been given to spatial heterogeneity, which is one of the fundamental spatial properties. Spatial heterogeneity in compositional data generally refers to the inconsistent relationships between the relative ratios of each composition and the associated factors across geographical space. This variability can be investigated by estimating spatially varying coefficients (SVCs) at each location [8]. To date, the methodology and application have not been widely discussed.

To enrich this research area, this study proposes a Moran eigenvector-based SVC (MSVC) [3] framework to explore the spatial heterogeneity of compositional data. MSVC links the local variations to the global spatial process, providing interpretable explanations of SVCs. In addition, based on the linear regression framework, MSVC has the advantage of being extendable to accommodate the specific properties of compositional data.

2 Properties of compositional data

Compositional data including D positive components can be represented by a vector $\mathbf{y} = (y_1, \dots, y_D)$, where each component y_j describes only relative information (e.g., proportion or percentage) and all of them sum up to a constant. \mathbf{y} is defined on a simplex space \mathbb{S}^D as



© Zhan Peng and Ryo Inoue;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 59; pp. 59:1–59:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$$\mathbb{S}^D = \left\{ \mathbf{y} = (y_1, \dots, y_D) \mid y_j > 0, j = 1, \dots, D; \sum_j y_j = 1 \right\}. \quad (1)$$

The constant-sum of compositions leads to spurious correlation [1], which poses challenges to the use of traditional statistical methods with compositional data. A common solution to this problem is to adopt the isometric log-ratio (ILR) transformation [2], which maps compositions \mathbf{y} from the simplex space \mathbb{S}^D to ILR coordinates \mathbf{y}^* in the Euclidean space \mathbb{R}^{D-1} through $\mathbf{y}^* = \text{ilr}(\mathbf{y}) := \mathbf{V}' \ln(\mathbf{y})$. The inverse ILR transformation is $\mathbf{y} = \text{ilr}^{-1}(\mathbf{y}^*) = \mathcal{C} \exp(\mathbf{V}\mathbf{y}^*)$, where \mathcal{C} is the closure operation that $\mathcal{C}\mathbf{y} := \mathbf{y} / \sum_j y_j$. The $D \times (D-1)$ matrix \mathbf{V} obeys $\mathbf{V}' \cdot \mathbf{V} = \mathbf{I}_{D-1}$ and $\mathbf{V} \cdot \mathbf{V}' = \mathbf{I}_D - (1/D)\mathbf{1}_{D \times D}$. Columns \mathbf{v}_i and vectors $\mathbf{e}_i = \mathcal{C} \exp(\mathbf{v}_i)$ forms orthonormal bases of \mathbb{R}^{D-1} and \mathbb{S}^D , respectively. The orthogonality of ILR coordinates allows for the use of classical regression models for each coordinate separately.

3 Method

3.1 MSVC model

The MSVC model is developed based on the correlation between eigenvalues and Moran's I statistic (MC). First, a spatial weight matrix \mathbf{C} is constructed by the binary relationships or distance decaying function (e.g., the exponential function). The eigenvector decomposition $(\mathbf{I} - \mathbf{1}\mathbf{1}'/N)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}'/N) = \mathbf{E}_N \mathbf{\Lambda} \mathbf{E}_N'$, where the left-hand side of the equation is also a part of MC, decomposes the spatial structure of the data into a set of orthogonal spatial patterns that are represented by each eigenvector in \mathbf{E}_N . $\mathbf{\Lambda}$ includes the corresponding eigenvalues.

Based on this work, Griffith (2008) [3] introduced a subset of eigenvectors into the basic linear model to account for the spatial heterogeneity in the regressed relationships. The resulting MSVC model is expressed as

$$\mathbf{y} = \sum_{k=0}^K \mathbf{x}_k \circ \beta_k^{ESF} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (2)$$

Here, $\beta_k^{ESF} = \beta_k \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_k$ represents the k -th spatially varying coefficient, which consists of the global trend of the spatial process $\beta_k \mathbf{1}$, and the linear combination of eigenvectors $\mathbf{E}\boldsymbol{\gamma}_k$ that account for the local deviations from the trend at each location. "o" is the column-wise product operator. The next section will extend the MSVC model to accommodate compositional data.

3.2 MSVC model for compositional data

Let $\mathbf{Y} = (\mathbf{y}_1 \ \dots \ \mathbf{y}_N)'$ represent N samples of D -composition data, where \mathbf{y}_i , $i = 1, \dots, N$, is the $D \times 1$ transposed vector of the i -th sample, and $\mathbf{y}_{(j)}$, $j = 1, \dots, D$ is the $N \times 1$ the vector of the j -th component. The ILR transformation of \mathbf{Y} becomes $\text{ilr}(\mathbf{Y}) = (\text{ilr}(\mathbf{y}_1) \ \dots \ \text{ilr}(\mathbf{y}_N))'$, where $\text{ilr}(\mathbf{y}_i) = \mathbf{y}_i^* = (y_{i(1)}^* \ \dots \ y_{i(D-1)}^*)$.

The MSVC model for the j -th ($j = 1, \dots, D-1$) coordinate is formulated as

$$\mathbf{y}_{(j)}^* = \sum_{k=0}^K \mathbf{x}_k \circ \left(\beta_{k(j)}^* \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_{k(j)}^* \right) + \boldsymbol{\varepsilon}_{(j)}^*, \quad \boldsymbol{\varepsilon}_{(j)}^* \sim \mathcal{N}(\mathbf{0}, \sigma_{(j)}^2 \mathbf{I}). \quad (3)$$

where $*$ denotes the ILR transformation, $\mathbf{x}_k (k = 0, \dots, K, \mathbf{x}_0 = \mathbf{1})$ is the k -th covariate, $\beta_{k(j)}^{SVC*} = \beta_{k(j)}^* \mathbf{1} + \mathbf{E} \gamma_{k(j)}^*$ represents the relationship between the k -th covariate and the j -th ILR coordinate. We can also rewrite the model into a more general form as

$$\mathbf{y}_{(j)}^* = \mathbf{X} \beta_{k(j)}^* + \tilde{\mathbf{E}} \gamma_{k(j)}^* + \varepsilon_{(j)}^*, \tag{4}$$

where $\tilde{\mathbf{E}} = (\mathbf{x}_0 \circ \mathbf{E}, \mathbf{x}_1 \circ \mathbf{E}, \dots, \mathbf{x}_K \circ \mathbf{E})$ are considered as proxy variables. Under the ILR transformation, Equation (4) can be estimated by ordinary linear regression for each $\mathbf{y}_{(j)}^*$, but the interpretation of the estimated coefficients is not straightforward. In line with [4, 8], we adopt the concept of semi-elasticity (SE), which reflects the relative percentage change in a particular composition with respect to a unit change in the covariate of interest. The k -th spatially varying SE of the j -th composition at the i -th location is defined as

$$e(y_j, \mathbf{x}_k)_i = \left(\ln \beta_{ik(j)} - \sum_{m=1}^D y_{i(m)} \ln \beta_{ik(m)} \right) y_{i(j)}. \tag{5}$$

where $y_j, y_{i(m)}, y_{i(j)}$, and $\beta_{ik(j)}$ are the inverse transformed variables in the simplex space.

3.3 Variable selection

Using all eigenvectors can result in an excessive number of explanatory variables. This can create computational challenges and potential overfitting problems. To mitigate these issues, as suggested by [7], we first select eigenvectors whose corresponding eigenvalues satisfy $\lambda_l / \lambda_{max} > 0.25^1$ and then use penalized regression (see Equation (6)) to choose only the eigenvectors that explain significant spatial variations in the data.

$$\min(\mathbf{y}_{(j)}^* - \mathbf{X} \beta_{k(j)}^* - \tilde{\mathbf{E}} \gamma_{k(j)}^*)' (\mathbf{y}_{(j)}^* - \mathbf{X} \beta_{k(j)}^* - \tilde{\mathbf{E}} \gamma_{k(j)}^*) + \lambda |\gamma_{k(j)}^*|_1 \tag{6}$$

The value of λ is determined by cross-validation or information criteria. Because the output of the penalized regression is known to be biased, we use it only for variable selection and apply the proposed model to estimate the coefficients of the selected variables.

4 Empirical application

4.1 Data and methods

We applied the proposed model to the analysis of the municipal-level household income structure of Tokyo, Japan in 2018. The annual income data were aggregated into three main groups: Low (less than 2 million JPY), Middle (between 2 and 7 million JPY), and High (more than 7 million JPY), resulting in a three-composition response variable. The following matrix \mathbf{V} for the ILR transformation of compositions generates two ILR coordinates [6].

$$\mathbf{V} = \begin{bmatrix} 2/\sqrt{6} & 0 \\ 1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}. \tag{7}$$

The first coordinate $\mathbf{y}_{(1)}^*$ refers to the relative importance of the low-income with respect to the other two groups, and the second coordinate $\mathbf{y}_{(2)}^*$ refers to that of the middle-income with respect to the high-income group.

¹ $0.25\lambda_{max}$ relates to roughly 5% of the variance in response variable attributable to positive spatial dependence.

The covariates used in the analysis included the proportion of people with secondary education (Uni), the unemployment rate (Unemp), the proportion of people aged over 65 (Age), and the homeownership rate (House). The data were published by the Statistics Bureau of Japan on the e-Stat portal site (<https://www.e-stat.go.jp/en>). We excluded 11 municipalities with no records, resulting in a final sample size of $N = 51$. Based on the adjacency of regions, we built a spatial weight matrix in which the (i, j) -th element was 1 if two regions i, j shared a common boundary, and 0 otherwise. From this matrix, we extracted 12 out of 51 eigenvectors to be further selected by the penalized regression.

4.2 Results and discussion

First, we conducted the ordinary linear regression without considering the spatial effects. The results shown in Table 1 suggest that all covariates except the unemployment rate are significantly associated with both ILR coordinates. The residual MC indicates that the spatial autocorrelation is significant in $\mathbf{y}_{(1)}^*$, but not significant in $\mathbf{y}_{(2)}^*$.

The results of the proposed model are summarized in Table 2. For $\mathbf{y}_{(1)}^*$, the use of eigenvectors led to a decrease in the residual MC and a noticeable increase in the adjusted R^2 , suggesting that the spatial variations captured by the eigenvectors explain a considerable proportion of the variance in the response variable. No eigenvector was found to be significant on $\mathbf{y}_{(2)}^*$, which aligns with the MC of $\mathbf{y}_{(2)}^*$ shown in Table 1 and proves that the proposed model can distinguish the existence of spatial heterogeneity. This result only indicates that the impacts of covariates on the ratio between middle- and high-income are spatially invariant. However, it does not necessarily imply that the impacts on each income group remain constant. For further analyzing their relationships, we can transform coefficients back to the simplex plane and then calculate the corresponding SEs (Equation (5)).

Figure 1 plots the SEs of each covariate across different income groups. The SEs provide insights into the interconnections among income groups, as they sum up to zero within each region for each covariate. For the entire region, we observe that an increase in the proportion of individuals with secondary education contributes to the shift from low- and middle-income to high-income groups. However, this impact varies by region. Particularly in the southeastern area, which serves as the business and cultural center of Tokyo, the expansion of the high-income group is notably significant. This can be attributed to the concentration of knowledge-intensive industries in this region, which has led to a higher demand for skilled professionals. In Chiyoda-ku, for example, when the proportion of the educated population increases by one unit, the high-income group increases by 0.426%, which

■ **Table 1** Estimation results of the ordinary linear regression.

Variables	$\mathbf{y}_{(1)}^*$		$\mathbf{y}_{(2)}^*$	
	Coefficient	Std. Error	Coefficient	Std. Error
Constant	-0.798*	0.465	0.673**	0.274
Uni	-0.678*	0.381	-1.159***	0.224
Unemp	0.110**	0.049	0.044	0.029
Age	2.967***	1.065	2.988***	0.626
House	-1.460***	0.351	-0.807***	0.206
MC	0.236***		0.014	
Adjusted R^2	0.568		0.860	

Note) : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

■ **Table 2** Estimation results of the MSVC-based regression.

Variables	$\mathbf{Y}_{(1)}^*$			$\mathbf{Y}_{(2)}^*$		
	Min.	Med.	Max.	Min.	Med.	Max.
Constant	-0.296	-0.239	-0.168		0.673	
Uni	-1.285	-1.048	-0.801		1.159	
Unemp		0.038			0.044	
Age		3.069			2.988	
House	-1.969	-1.757	-1.547		-0.807	
MC		-0.005**			0.014	
Adjusted R^2		0.729			0.860	

Note) : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

is the highest among all regions. The middle- and low-income groups decrease by 0.153% and 0.273%, respectively. In contrast, in Hinode-machi, which is located on the periphery of Tokyo, the high-income group increases by only 0.189%, and the low- and middle-income groups decrease by only 0.094% and 0.096%, respectively. An increase in the unemployment rate results in the expansion of low- and middle-income groups, along with a decrease in the proportion of the high-income group, primarily observed in southeastern Tokyo. The proportion of people aged over 65 negatively affects the high-income group but positively affects the other two groups. This is consistent with the fact that older people generally have lower incomes and may require more social welfare support. Furthermore, this impact is stronger compared to other factors in terms of the magnitude of the SE, highlighting the importance of considering the impact of the aging of population on income analysis. Lastly, the increase in homeownership rate contributes to the transition of low-income groups into middle-income groups in western and northeastern Tokyo. The middle-income group further shifts to high-income in the southeastern parts.

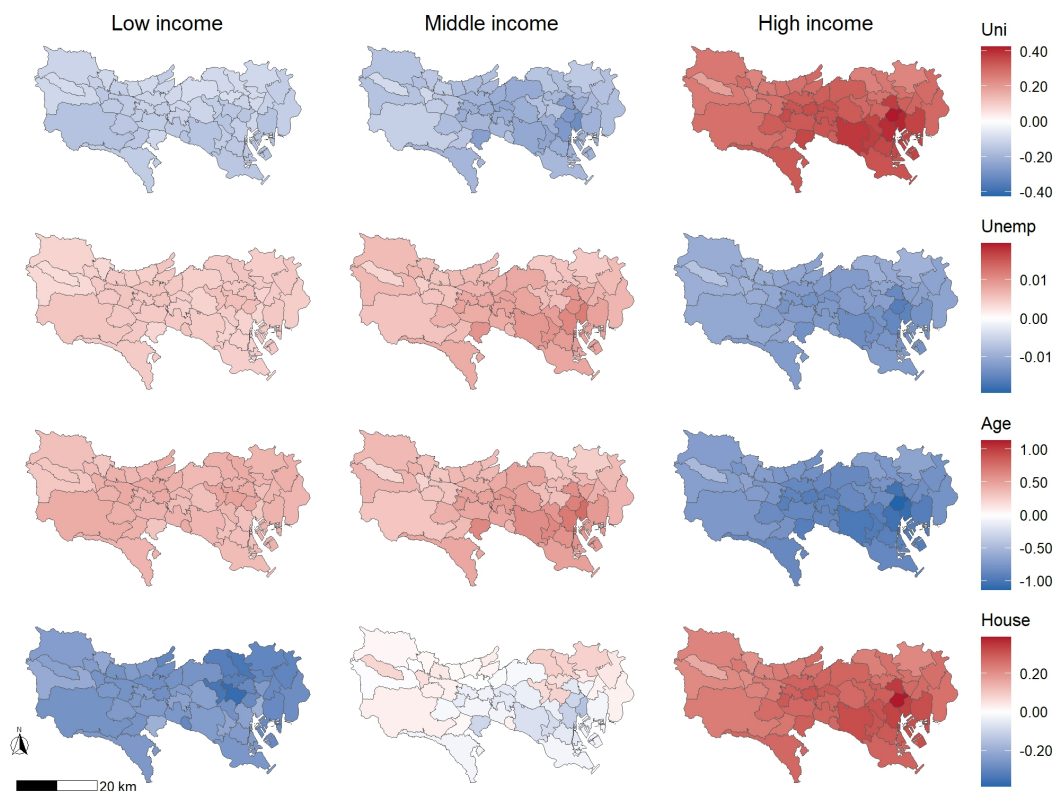
5 Conclusion

This study proposed an MSVC-based framework to investigate the spatial heterogeneity of compositional data. We adopted the ILR transformation and the semi-elasticity to aid the model estimation and interpretation. The application on household income in Tokyo indicated that socio-economic factors affect income distribution differently across regions, which yields insights for understanding the drivers of income inequality.

There are still many challenges and our work is only just beginning. It is worth discussing in the future a more intuitive way of model interpretation. Moreover, an in-depth investigation is necessary to assess the impact a change in the type of spatial weights matrix and the criteria for selecting eigenvectors might have on the outputs. Finally, comparing the performance of the proposed method and previous approaches in analysing spatial heterogeneity would be an interesting topic for future discussions.

References

- 1 J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., GBR, 1986.
- 2 J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, April 2003. doi:10.1023/A:1023818214614.



■ **Figure 1** Spatial distribution of semi-elasticities of MSVC-based CoDA.

- 3 Daniel A Griffith. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A: Economy and Space*, 40(11):2751–2769, November 2008. doi:10.1068/a38218.
- 4 Joanna Morais, Christine Thomas-Agnan, and Michel Simioni. Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47(5):1–25, September 2018. doi:10.17713/ajs.v47i5.718.
- 5 Vera Pawlowsky-Glahn and Juan José Egozcue. Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration*, 164:28–32, May 2016. doi:10.1016/j.gexplo.2015.12.010.
- 6 Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modelling and Analysis of Compositional Data*. John Wiley & Sons, 2015. doi:10.1002/9781119003144.
- 7 Hajime Seya, Daisuke Murakami, Morito Tsutsumi, and Yoshiki Yamagata. Application of LASSO to the eigenvector selection problem in eigenvector-based spatial filtering. *Geographical Analysis*, 47(3):284–299, 2015. doi:10.1111/gean.12054.
- 8 Takahiro Yoshida, Daisuke Murakami, Hajime Seya, Narumasa Tsutsumida, and Tomoki Nakaya. Geographically weighted regression for compositional data: An application to the U.S. household income compositions. *GIScience 2021 Short Paper Proceedings. 11th International Conference on Geographic Information Science. September 27-30, 2021*. Poznań:Poland (Online), 2021. doi:10.25436/E2G599.