

組成データのための地理的加重回帰モデル

吉田崇紘*・村上大輔**・瀬谷創***・
堤田成政****・中谷友樹*****・堤盛人*****

Geographically Weighted Regression for Compositional Data

Takahiro YOSHIDA*, Daisuke MURAKAMI**, Hajime SEYA***,
Narumasa TSUTSUMIDA****, Tomoki NAKAYA*****, Morito TSUTSUMI*****

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. In this study, we propose a GWR model for compositional data, which we term GWRCoDa. It allows us to model spatially varying relationships with considering the constant-sum constraint. We apply this model to analyze household income compositions at the county-level in the US. The spatial varying compositional semi-elasticities are also proposed, and its interpretational usefulness is empirically suggested.

Keywords: 組成データ (compositional data), 定数和制約 (constant-sum constraint), 対数比変換 (log-ratio transformation), 空間的異質性 (spatial heterogeneity), 地理的加重回帰モデル (geographically weighted regression)

1. はじめに

地理的加重回帰 (geographically weighted regression: GWR; Fotheringham et al., 1996; Brunsdon et al., 1996) は、場所毎に回帰係数を推定することで各地点の局所的な特性を捉えようという手法である。その実用性・可用性の高さから、環境科学や生態学などの幅広い分野で応用されてきた。地理情報が多様化しつつある昨今、大規模なデータにおいても高速かつ安定して推定可能とするモデル (Murakami et al., 2020) や、ロジスティック分布 (Atkinson et al., 2003) やポアソン分布 (Nakaya et al., 2005)、負の二項分布 (da Silva and Rodrigues, 2014) などの正規分布以外の統計分布に従うデータに対応したモデルの開発などを通じて、GWR の洗練と拡張が進められている。

本研究では、組成データ (compositional data: CoDa) と呼ばれる、所得階層別や年齢階級別の人口構成比

や罪種別の犯罪率のような、要素の値が比率や割合で表される多変量データの GWR を検討する。da Silva and Lima (2017) は二変量の比率データに対する GWR として、ベータ分布を導入したモデルを提案している。多変量の比率データである CoDa に対する GWR は、da Silva and Lima (2017) からの自然な拡張として、ベータ分布の一般形であるディリクレ分布を導入する方法に加え、岩石の化学組成データなどを扱う地質学分野において発展し、対数比変換に基づいた組成データ解析 (Compositional data analysis: CoDA; Aitchison, 1982, 1986) の手法を導入する方法が考えられる。

両方法の大きな違いは、CoDa を統計分析する事前処理として、変数変換を行うかどうかにある。組成データは、各要素の総和が一定 (たとえば、割合であれば 1, 百分率であれば 100) という定数和制約

* 正会員 国立環境研究所 (National Institute for Environmental Studies)
〒305-8506 茨城県つくば市小野川 16-2 E-mail: yoshida.takahiro@nies.go.jp
** 正会員 統計数理研究所 (The Institute of Statistical Mathematics)
*** 正会員 神戸大学 (Kobe University)
**** 正会員 京都大学 (Kyoto University)
***** 正会員 東北大学 (Tohoku University)
***** 正会員 筑波大学 (University of Tsukuba)

(constant-sum constraint) により、その標本空間は単体空間上に限定される。前者の方法は、ディリクレ分布を仮定し、単体空間上のデータを直接モデル化する。一方、後者の方法は、対数比変換により組成データを単体空間からユークリッド空間に写像しモデル化する。前者の方法では、組成データの各要素間の相関が常に負であるなどのディリクレ分布に仮定される条件をおく必要があり、定数和制約により二変量の場合は必然的に満たされるこの条件が、三変量以上の多変量においては満たされないことも多い。一方、後者の方法は、対数比変換によりパラメータの解釈が直接的でなくなる欠点がある。しかしながら、近年の CoDA における対数比変換の詳細な検討から、変換後のデータにおいても等長性 (isometry) などの変量間の関係を担保しつつ、統計解析後の逆変換を行うことで容易な解釈を可能とする方法が提案されるなど、欠点は解消されつつある (Pawlowsky-Glahn et al., 2015)。

以上を踏まえ本研究では、CoDA の方法による組成データのための GWR (GWRCODA) を検討する。そして、開発したモデルを米国の所得階層別の世帯比率の分析に適用することでその有効性を確認する。

2. 組成データ解析

2.1. 基礎演算子と距離関数

本節では、GWRCODA を導入する準備として、CoDA における基礎演算子と距離関数、対数比変換を整理する。

変量数が D (D -parts), 各変量が比率で与えられた組成データベクトル $\mathbf{p} = (p_1, p_2, \dots, p_D)$ の標本空間 (単体空間) は式(1)で定義することができる：

$$\mathcal{S}^D = \left\{ \mathbf{p} \mid p_m > 0, m = 1, 2, \dots, D, \sum_{m=1}^D p_m = 100 \right\}. \quad (1)$$

CoDA において、 $\mathcal{S}^D$ における $(D-1)$ 次元のベクトル空間は、摂動演算 (perturbation operation, 式 2)：

$$\mathbf{p} \oplus \mathbf{q} = \mathcal{C}(p_1 \cdot q_1, p_2 \cdot q_2, \dots, p_D \cdot q_D) \in \mathcal{S}^D \quad (2)$$

と、冪演算 (powering operation, 式 3)：

$$\alpha \odot \mathbf{p} = \mathcal{C}(p_1^\alpha, p_2^\alpha, \dots, p_D^\alpha) \in \mathcal{S}^D \quad (3)$$

によって定義される (Aitchison, 1986)。ここで、 $\mathbf{p}, \mathbf{q} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$ であり、 $\mathcal{C}(\cdot)$ は式(4)で表される閉鎖演算 (closure operation)：

$$\mathcal{C}(\mathbf{z}) = \left(\frac{100 \cdot z_1}{\sum_{m=1}^D z_m}, \frac{100 \cdot z_2}{\sum_{m=1}^D z_m}, \dots, \frac{100 \cdot z_D}{\sum_{m=1}^D z_m} \right) \in \mathcal{S}^D \quad (4)$$

である。ここで、 $\mathbf{z} = (z_1, z_2, \dots, z_D) \in \mathbb{R}_+^D$ である。摂動演算、冪演算はそれぞれユークリッド空間上の足し算、掛け算に相当する。なお、線形結合は式(5)のように表すことができる：

$$\begin{aligned} & \bigoplus_{k=1}^K \alpha_k \odot \mathbf{p}_k \\ &= \alpha_1 \odot \mathbf{p}_1 \oplus \alpha_2 \odot \mathbf{p}_2 \oplus \dots \oplus \alpha_K \odot \mathbf{p}_K \quad (5) \\ &= \mathcal{C} \left(\prod_{k=1}^K p_{k,1}^{\alpha_k}, \prod_{k=1}^K p_{k,2}^{\alpha_k}, \dots, \prod_{k=1}^K p_{k,D}^{\alpha_k} \right) \in \mathcal{S}^D. \end{aligned}$$

ここで、 $\mathbf{p}_k \in \mathcal{S}^D$, $\alpha_k \in \mathbb{R}$, $k = 1, 2, \dots, K$ である。

また、Aitchison (1992) は、 $\mathcal{S}^D$ における距離関数を議論し、アイチソン内積 (Aitchison inner product, 式 6)：

$$\langle \mathbf{p}, \mathbf{q} \rangle_A = \frac{1}{2D} \sum_{m=1}^D \sum_{\tilde{m}=1}^D \ln \frac{p_m}{p_{\tilde{m}}} \ln \frac{q_m}{q_{\tilde{m}}} \quad (6)$$

とアイチソンノルム (Aitchison norm, 式 7)：

$$\|\mathbf{p}\|_A = \sqrt{\langle \mathbf{p}, \mathbf{p} \rangle_A} = \sqrt{\frac{1}{2D} \sum_{m=1}^D \sum_{\tilde{m}=1}^D \left(\ln \frac{p_m}{p_{\tilde{m}}} \right)^2} \quad (7)$$

を定義している。ここで、 $m, \tilde{m} = 1, 2, \dots, D$ である。式(6), (7)から確認できるように、 $\mathcal{S}^D$ における距離関数はユークリッド空間のものと比較して、対数比に基づいて定義される。

2.2. 対数比変換

CoDA における分析手順は、対数比変換 (単体空間)、統計解析 (ユークリッド空間)、逆対数比変換 (単体空間)、そして解釈、とすることが多い。標準的に用いられる統計解析手法がユークリッド空間を前提に定義されているため、組成データをその標本空間である単体空間からユークリッド空間上に変換することで豊富な解析手段を享受することができる。このための代表的な対数比変換法として、相加対数比 (additive log-ratio: alr, 式 8)：

$$\text{alr}(\mathbf{p}) = \left(\ln \frac{p_1}{p_D}, \ln \frac{p_2}{p_D}, \dots, \ln \frac{p_{D-1}}{p_D} \right) \in \mathbb{R}^{D-1}, \quad (8)$$

有心対数比 (centered log-ratio: clr, 式 9) :

$$\text{clr}(\mathbf{p}) = \left(\ln \frac{p_1}{h(\mathbf{p})}, \ln \frac{p_2}{h(\mathbf{p})}, \dots, \ln \frac{p_D}{h(\mathbf{p})} \right) \in \mathbb{R}^D, \quad (9)$$

そして、等長対数比 (isometric log-ratio, ilr, 式 10) :

$$\text{ilr}(\mathbf{p}) = \mathbf{p}^* = (p_1^*, p_2^*, \dots, p_{D-1}^*) \in \mathbb{R}^{D-1} \quad (10)$$

がある (Pawlowsky-Glahn et al., 2015). ここで、 $h(\mathbf{p}) = (\prod_{m=1}^D p_m)^{1/D}$ は幾何平均であり、 $p_l^* = \sqrt{\frac{D-l}{D-l+1}} \ln \frac{p_l}{\sqrt{\prod_{m=l+1}^D p_m}}$, $l = 1, 2, \dots, (D-1)$ である.

alr 変換は、ロジット変換の一種としてよく知られており、特に交通工学の分野における集計型ロジット (aggregated logit) モデルの被説明変数として用いられている. clr 変換は、マーケティングの分野における積乗型競合相互作用 (multiplicative competitive interaction: MCI) モデルに用いられている. ilr 変換は、等長性を保つ望ましい対数比として Egozcue et al. (2003) において CoDA の分野の中で提案された変換法である. clr 変換と ilr 変換は、等長性 (式 11) :

$$\langle \mathbf{p}, \mathbf{q} \rangle_A = \langle \text{clr}(\mathbf{p}), \text{clr}(\mathbf{q}) \rangle = \langle \text{ilr}(\mathbf{p}), \text{ilr}(\mathbf{q}) \rangle, \quad (11)$$

$$\| \mathbf{p} \|_A = \| \text{clr}(\mathbf{p}) \| = \| \text{ilr}(\mathbf{p}) \|,$$

を満たすが、alr 変換は満たさない. clr 変換は、幾何平均で除することにより、変換した要素の和がゼロとなり、共分散行列が特異となる. ilr 変換は、他変換法で生じる問題を避け、任意の正規直交座標のもとに変換を行うことが可能である. このため様々な ilr 変換が定義できるが、式(10)で定義した ilr 変換は、第一変量 (pivot) のみを変換後も直接解釈しやすい変換である (Fišerová and Hron, 2011). なお、ilr 変換の正規直交基底の取り方は、順・逆対数比変換において共通であれば解釈に影響しないことが知られている (Filzmoser et al., 2018).

3. 地理的加重回帰モデル

GWR は地点 s_i ($i \in \{1, 2, \dots, n\}$) の被説明変数 y_i が式(12)に従うことを仮定する :

$$y_i = \sum_{k=1}^K x_{i,k} \beta_{i,k} + \varepsilon_i \quad (12)$$

ここで、 $x_{i,k}$ は説明変数、 ε_i は誤差項である. 地点 s_i 周辺の局所的な特性 (空間的異質性: spatial heterogeneity) を捉えるために、回帰係数は地点 s_i からの距離に応じて各標本に重みを付けた上で推定される. その推定量は式(13)で表される :

$$\hat{\beta}_i = (\mathbf{X}' \mathbf{G}_i \mathbf{X})^{-1} \mathbf{X}' \mathbf{G}_i \mathbf{y} \quad (13)$$

ここで、 $\mathbf{y}$ は被説明変数ベクトル、 $\mathbf{X}$ は説明変数行列である. $\mathbf{G}_i$ は地点 s_j ($j \in \{1, 2, \dots, n\}$) に対する重み $g_{i,j}$ を第 j 要素に持つ対角行列である. この重みは、例えば正規カーネル (式 14) で与えることができる :

$$g_{i,j} = \exp \left[- \left(\frac{d_{i,j}}{b} \right)^2 \right] \quad (14)$$

ここで、 $d_{i,j}$ は地点 s_i, s_j 間の直線距離である. 回帰係数はバンド幅 b が大きい場合に大域的、小さい場合に局所的な空間パターンを示す.

b は Leave-One-Out Cross-Validation (LOOCV) によって推定することができる. LOOCV では「 i 以外の標本の値から標本 i の値を予測する」操作を全標本について実行し、それを反復することで CV スコア (式 15) を最小化する b を探索する.

$$\text{CV} = \sum_{i=1}^n \left[y_i - \sum_{k=1}^K x_{i,k} \hat{\beta}_{-i,k} \right]^2 \quad (15)$$

ここで、 $\hat{\beta}_{-i,k}$ は i 以外の標本から推定される $\beta_{i,k}$ であり、 $\mathbf{G}_i$ の第 i 要素を 0 に置き換えた式 (13) で与えられる.

4. 組成データのための地理的加重回帰モデル

本節では、前節までに整理した CoDA と GWR を組み合わせた GWRCoDa を検討する.

式(12)において、被説明変数が D -parts の CoDa の場合を考える. このとき、 $\mathcal{S}^D$ における GWR (GWRCoDa) は式(16)で表すことができる :

$$\mathbf{y}_i = \bigoplus_{k=1}^{K+1} (x_{i,k} \odot (\boldsymbol{\beta}_i)_k) \oplus \boldsymbol{\varepsilon}_i, \quad (16)$$

ここで、 $\mathbf{y}_i$, $(\boldsymbol{\beta}_i)_k$, $\boldsymbol{\varepsilon}_i \in \mathcal{S}^D$, $x_{i,k} \in \mathbb{R}$ である. GWRCoDa のパラメータを推定するためには、定数と制約と空間的異質性を考慮する必要がある.

まず、定数和制約を考慮するため、式(12)の両辺に

ilr 変換 (式 10) を行う. ilr 変換後の GWRCoDa は式(17)で表すことができる.

$$y_i^{*(l)} = \sum_{k=1}^{K+1} \left(x_{i,k} \cdot (\beta_i^{*(l)})_k \right) + \varepsilon_i^{*(l)}, \quad (17)$$

ここで, 添え字「 (l) 」は $(D-1)$ 本ある回帰モデルの第 l 番目であることを表す.

続いて, 空間異質性を考慮する. ilr 変換の等長性より, 式(16)の残差 $\mathbf{u}_i \in \mathcal{S}^D$ の二乗和は式(18)で表すことができることから, この最小化は $(D-1)$ 本の回帰モデルの残差を個別に最小化した場合に等しくなる. したがって, 各回帰モデルを独立に扱うことが可能となり, 第3節で整理した GWR および LOOCV を独立に $(D-1)$ 回実行することでパラメータ推定値を得ることができる.

$$\begin{aligned} & \sum_{i=1}^n \|\mathbf{u}_i\|_A^2 \\ &= \sum_{i=1}^n \left\| \mathbf{y}_i \oplus \left[(-1) \odot \bigoplus_{k=1}^{K+1} (x_{i,k} \odot (\beta_i)_k) \right] \right\|_A^2 \quad (18) \\ &= \sum_{i=1}^n \|\mathbf{u}_i^*\|^2 = \sum_{i=1}^n \sum_{l=1}^{D-1} (\varepsilon_i^{*(l)})^2. \end{aligned}$$

第 l 番目のパラメータとバンド幅は, 式(13), (15)と同様に得ることができる (式 19, 20) :

$$\hat{\beta}_i^{*(l)} = [\mathbf{X}'\mathbf{G}_i(b^{(l)})\mathbf{X}]^{-1} \mathbf{X}'\mathbf{G}_i(b^{(l)})\mathbf{y}^{*(l)}, \quad (19)$$

$$CV = \sum_{i=1}^n \left(y_i^{*(l)} - \sum_{k=1}^{K+1} \left(x_{i,k} \cdot (\hat{\beta}_i^{*(l)})_k \right) \right)^2. \quad (20)$$

以上より, ユークリッド空間における GWR のパラメータと残差が得られる. これらに逆対数比変換(式 10)を行うことで, 説明変数の偏弾力性 (式 21) と予測確率 $\hat{\mathbf{y}}_i$ を導出することができる (Egozcue et al., 2011, Morais et al., 2018).

Semi elasticity $_{i,m}$

$$= \left(\ln((\beta_i)_{k,m}) - \sum_{m=1}^D \left(y_{i,m} \ln((\beta_i)_{k,m}) \right) \right) y_{i,m}. \quad (21)$$

5. 所得分析への応用

5.1. 概要

本節では GWRCoDa を米国の郡単位の世帯所得データ (2017 年) に適用する. 分析では, 地理的な連担性を保つためアラスカ州とハワイ州を除いた. サンプルサイズは 3,108 である. 被説明変数 (CoDa) は, 過去 12 カ月の世帯所得に応じて, 75,000 ドル以上 (High-Income), 35,000 ドルから 75,000 ドルの間 (Middle-Income), 35,000 ドル以下 (Low-Income) の 3 変量の世帯数比率とした. 図 1 は組成データの 3 次元単体空間上の分布である. 地域的な空間異質性を考察するため, 米国内有数の都市部である New York 郡 (ニューヨーク州), メキシコ国境に位置する El Paso 郡 (テキサス州), 国立公園内に位置し自然観光が盛んな Park 郡 (ワイオミング州) を取り上げる (図 2). 説明変数には, 学位取得者・大卒者の比率 (Univ.), 年齢の中央値 (Age), および英語が主要言語である居住者の比率 (Eng.) とした. 各変数は, 統計ソフト R の tidyrcensus パッケージから収集した.

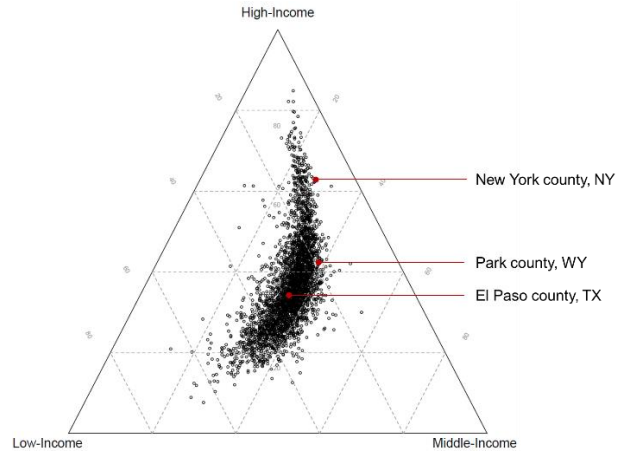


図 1: 所得階層別世帯数割合の三角図

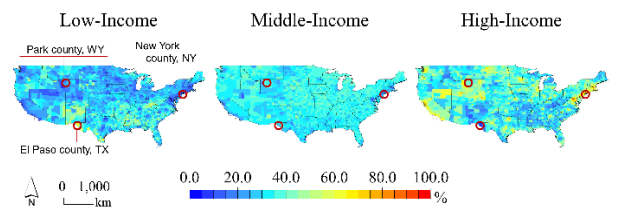


図 2: 所得階層別世帯数割合の空間分布

5.2. 結果

推定された各偏弾力性を図3に示す。偏弾力性は、説明変数が1単位増加したときの被説明変数の相対的变化率を示す。興味深いことに、CoDaの回帰モデルにおいては、説明変数毎の偏弾力性の和が0であるため、各変数の影響を容易かつ直接的に解釈することができる。たとえば、都市部であるNew York郡のUniv.が1単位増加した場合、高所得比率は+0.513%、中所得比率は-0.256%、低所得比率は-0.257%となった。同様に、El Paso郡では、高所得比率は+0.509%、中所得比率は-0.066%、低所得比率は-0.443%であり、Park郡では、高所得比率は+0.282%、中所得比率は-0.151%、低所得比率は-0.131%であった。New York郡の低所得者比率に対するUniv.の影響は、Park郡と比べ約2倍であった。図3の偏弾力性の空間分布からは、特に東海岸と西海岸において、Univ.が高所得者比率に正の影響を与えることを示している。これらの地域には多くのホワイトカラーと専門労働者が住んでいるため、この結果は妥当といえる。

図4は、各地域において、説明変数の値を変化させたときに世帯数比率がどの程度変化するかを予測確率によって示したものである。図より、いずれの地域においても、Univ.は高所得比率に正の影響をもつことがわかる。特にNew York郡において強い正の関係があり、他変数が一定と仮定した場合、Univ.が8%から10%程度となると支配的な所得者層が変化する。AgeとEng.は地域ごとに異なる変化パターンを示す。New York郡ではAgeの各所得比率への影響は10%未満でほぼ変わらないのに対し、Ageが増加すると、El Paso郡では低所得者比率が線形に増加し、Park郡では指数的に増加する。この結果は、Park郡における低所得比率へのAgeの影響が他地域に比べてより重要であることを示唆している。同様に、New York郡におけるEng.の各所得比率への影響はほぼ変わらないのに対し、El Paso、Parkの両郡では高所得比率が増加する。また、Park郡では、低所得者比率が指数的に減少している。この結果は、Park郡では、低所得者比率を下げ、高所得者比率を上げるためにEng.が重要であることを示唆する。

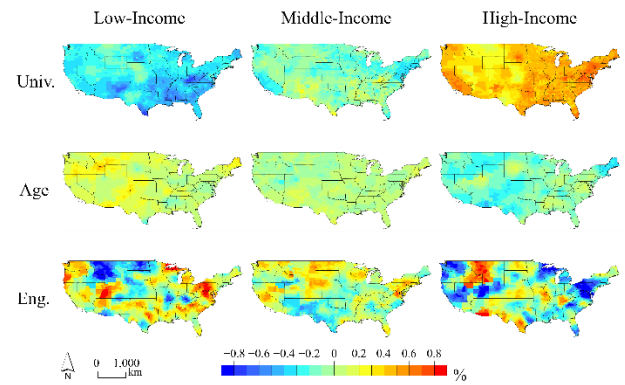


図3: 偏弾力性の空間分布

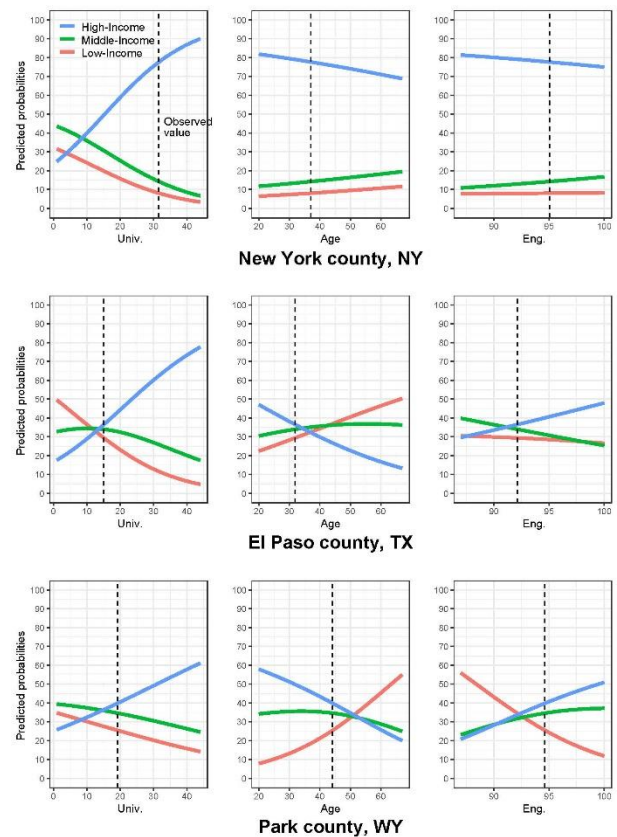


図4: 予測確率の変化 (各パネルにおいて、対象の説明変数は全観測値の[最小値, 最大値]の範囲で変化させ、その他の説明変数は各郡の観測値で固定)

6. おわりに

本研究では、多変量の比率からなる組成データのためのGWRとして、CoDaの対数比変換法を採用し、定数和制約と空間異質性を考慮可能なGWRCODAを開発した。それを米国の所得階層別の

世帯比率の分析に適用した。その結果から、説明変数の偏弾力性を解釈のしやすい形で空間パターンとともに示すことができ、また予測確率から地点別に説明変数が増減した場合の各所得比率の変化を示すことができた。

謝辞

本研究にあたり、ROIS-DS-JOINT (003RP2020)、JSPS 科研費 (17K18554, 18H03628) の助成を受けた。

参考文献

- Aitchison, J. (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44** (2), 139–160.
- Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.
- Aitchison, J. (1992) On criteria for measures of compositional difference. *Mathematical Geology*, **24**(4), 365–379.
- Atkinson, P.M., German, S.E., Sear, D.A., and Clark, M.J. (2003) Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis*, **35** (1), 58–82.
- Brunsdon, C., Fotheringham, A.S., and Charlton, M.E. (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28** (4), 281–298.
- da Silva, A.R. and Lima, A.D.O. (2017) Geographically weighted beta regression. *Spatial Statistics*, **21**, 279–303.
- da Silva, A.R. and Rodrigues, T.C.V. (2013) Geographically weighted negative binomial regression—incorporating overdispersion. *Statistics and Computing*, **24**, 769–783.
- Egozcue, J.J., Daunis-i-Estadella, P., Pawlowsky-Glahn, V., Hron, K., and Filzmoser, P. (2011) Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, **6** (1-2), 87–108.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35** (3), 279–300.
- Filzmoser, P., Hron, K., and Templ, M. (2018) *Applied Compositional Data Analysis*. Cham: Springer.
- Fišerová, E. and Hron, K. (2011) On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, **43** (4), 455–468.
- Fotheringham, A.S., Charlton, M., and Brunsdon, C. (1996) The geography of parameter space: an investigation of spatial non-stationarity. *International Journal of Geographical Information Systems*, **10** (5), 605–627.
- Morais, J., Thomas-Agnan, C., and Simioni, M. (2018) Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, **47** (5), 1–25.
- Murakami, D., Tsutsumida, N., Yoshida, T., Nakaya, T., and Lu, B. (2020) Scalable GWR: A linear-time algorithm for large-scale geographically weighted regression with polynomial kernels. *Annals of the American Association of Geographers*, in press.
- Nakaya, T., Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2005) Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, **24** (17), 2695–2717.
- Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, R. (2015) *Modelling and Analysis of Compositional Data*. Chichester, UK: John Wiley & Sons.