定数和制約と誤差相関を考慮した組成データのための地理的加重回帰

吉田崇紘*・村上大輔・瀬谷創***・堤田成政****・中谷友樹*****

Geographically Weighted Regression for Compositional Data Considering the Constant Sum-Constraint and Correlated Errors

Takahiro Yoshida*, Daisuke Murakami**, Hajime Seya***,
Narumasa Tsutsumida****, Tomoki Nakaya*****

Abstract: This study proposes geographically weighted seemingly unrelated regression for compositional data (GWSURcoda) by combining geographically weighted regression (GWR), seemingly unrelated regression (SUR), and compositional data analysis (CoDA). GWR allows the modeling of spatial heterogeneity in regression models and is increasingly used in various fields. SUR assumes correlated errors among regression models. CoDA provides unique and useful tools for compositional data, which are restricted by a constant-sum constraint. It allows us to model spatially varying relationships with considering correlated errors and the constant-sum constraint. We apply GWSURcoda to analyze household income compositions at the county-level in the US. The spatial varying compositional semi-elasticities showed insightful and easy interpretation to understand spatial heterogeneity.

Keywords: 地理的加重回帰（geographically weighted regression），空間的異質性（spatial heterogeneity），組成データ（compositional data），定数和制約（constant sum-constraint），誤差相関（correlated errors）

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，空間的異質性（spatial heterogeneity）を捉えるため地点毎あるいはゾーン毎に回帰係数を推定する局所回帰モデルである．多種多様な地理空間情報が利用可能になってきた中で，GWR は，大規模データに対しても安定かつ高速に推定可能とする方法（e.g., Murakami et al. 2021）が提案されたり，ロジスティック分布（Atkinson et al., 2003）やポアソン分布（Nakaya et al., 2005）など正規分布以外の統計分布に従うデータを扱えるよう拡張が図られたりと，研究が蓄積されてきた．

GWR を含む地理的加重法の研究動向と今後の展望をレビューした堤田ほか（2021）が指摘する研究余地のひとつに，比率データに対応した GWR の拡張がある．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への対応として，da Silva and Lima (2017) の考え方を一般化し，ディリクレ分布を導入した GWR を構築することが考えられるがこのモデルはまだ開発されていない．

一方，別のアプローチに基づいて，多変量の比率データに対する GWR が提案されている．吉田ほか（2020）は，多変量の比率データの扱いについて，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，対数比変換により比率データを分析する枠組みである組成データ解析（compositional data analysis: CoDA; Aitchison, 1982）を援用したモデルを検討している．より具体的には，比率データの特性である各要素の総和が一定（たとえば，割合であれ

*　正会員　東京大学 大学院工学系研究科（The University of Tokyo）
〒113-8656　東京都文京区本郷 7-3-1　E-mail：yoshida.takahiro@up.t.u-tokyo.ac.jp
**　正会員　統計数理研究所 データ科学研究系（The Institute of Statistical Mathematics）
***　正会員　神戸大学 大学院工学研究科（Kobe University）
****　正会員　埼玉大学 大学院理工学研究科（Saitama University）
*****　正会員　東北大学 大学院環境科学研究科（Tohoku University）

ば 1，百分率であれば 100）という定数和制約（constant-sum constraint）を取り除いたうえで，変換した各変量について GWR を実行し，逆対数比変換によりパラメータの解釈を行うモデルを検討している．しかし，上記の方法では，定数和制約と空間的異質性は考慮されているものの，変換した変量に関する複数本の GWR の誤差相関は考慮されていない．後述するように CoDA の回帰モデルでは，対数比変換後の各モデルを独立とみなして推定が実行されることが多い．対数比変換自体に比率データの変量間の相互依存関係があるため，変換後の複数本のモデルを同時方程式モデルとして扱うことにより誤差相関を考慮することが発展の一つとして考えられる．

以上を踏まえ本研究では，吉田ほか（2020）のモデルに，同時方程式モデルである seemingly unrelated regression (SUR; Zellner, 1962) を組み合わせて誤差相関を考慮した geographically weighted and seemingly unrelated regression for compositional data (GWSURcoda) 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．なお，以降，多変量の比率データを組成データ（compositional data）と呼称する．

２．地理的加重回帰モデル

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

$$y_i = \sum_{k=1}^{K} x_{i,k}\beta_{i,k} + \varepsilon_i. \tag{1}$$

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（2）で表される：

$$\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{G}_i\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}_i\mathbf{y}. \tag{2}$$

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，例えば式（3）の正規カーネルで与えることができる：

$$g_{i,j} = \exp\left[-\left(\frac{d_{i,j}}{b}\right)^2\right]. \tag{3}$$

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．

バンド幅 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式（4）の CV スコア：

$$\mathrm{CV} = \sum_{i=1}^{n}\left[y_i - \sum_{k=1}^{K} x_{i,k}\hat{\beta}_{-i,k}\right]^2 \tag{4}$$

を最小化する b を探索する．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(2)で与えられる．次章では，GWR を組成データ用に拡張する．

３．組成データのための地理的加重回帰モデル

3.1. 組成データ

GWR の拡張に際し，まず組成データについて簡単に整理する．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（5）で定義することができる（Aitchison, 1982）：

$$\mathcal{S}^D = \left\{\mathbf{p}\,\middle|\, p_m > 0, m = 1,2,\dots,D, \sum_{m=1}^{D} p_m = 100\right\}. \tag{5}$$

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）：

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathrm{A}} = \frac{1}{2D}\sum_{m=1}^{D}\sum_{\dot{m}=1}^{D} \ln\frac{p_m}{p_{\dot{m}}}\ln\frac{q_m}{q_{\dot{m}}} \tag{6}$$

とアイチソンノルム（Aitchison norm，式 7）：

$$\| \mathbf{p} \|_{\mathrm{A}} = \sqrt{\langle \mathbf{p}, \mathbf{p} \rangle_{\mathrm{A}}} = \sqrt{\frac{1}{2D}\sum_{m=1}^{D}\sum_{\dot{m}=1}^{D}\left(\ln\frac{p_m}{p_{\dot{m}}}\right)^2} \tag{7}$$

を定義している．ここで，$\mathbf{p}, \mathbf{q} \in \mathcal{S}^D$，$m, \check{m}, \dot{m} = 1,2,\dots,D$ 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．

3.2. 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．様々な対数比変換法が検討されてきたが，ここでは現在よく利用される等長対数比（isometric log-ratio, ilr，式 8）変換を導入する．

$$\mathrm{ilr}(\mathbf{p}) = \mathbf{p}^* = (p_1^*, p_2^*, \dots, p_{D-1}^*) \in \mathbb{R}^{D-1} \qquad (8)$$

ここで, $p_l^* = \sqrt{\frac{D-l}{D-l+1}} \ln \frac{p_l}{\sqrt[D-l]{\Pi_{m=l+1}^{D} p_m}}$, $l = 1, 2, \dots, (D-1)$ である． ilr 変換は，等長性（式 9）を保つ望ましい対数比として Egozcue et al. (2003) が提案した変換法である．

$$\begin{aligned} \langle \mathbf{p}, \mathbf{q} \rangle_{\mathrm{A}} &= \langle \mathrm{ilr}(\mathbf{p}), \mathrm{ilr}(\mathbf{q}) \rangle, \\ \| \mathbf{p} \|_{\mathrm{A}} &= \| \mathrm{ilr}(\mathbf{p}) \|, \end{aligned} \qquad (9)$$

ilr 変換は，その他の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(8)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くした変換である（Fišerová and Hron, 2011）.

3.3. GWSURcoda

GWSURCoDa は，定数和制約，空間的異質性，誤差相関の 3 点の特性を考慮する．以下それぞれの対処について整理する．ここでは，被説明変数が D-parts の組成データの場合を考える.

まず，定数和制約を考慮するため，式（1）において，被説明変数$\mathbf{y}_i \in \mathcal{S}^D$，説明変数$x_{i,k} \in \mathbb{R}$，パラメータ$(\boldsymbol{\beta}_i)_k \in \mathcal{S}^D$，誤差項$\boldsymbol{\varepsilon}_i \in \mathcal{S}^D$に置き換え，ilr 変換すると式（10）のように表すことができる：

$$y_i^{*(l)} = \sum_{k=1}^{K+1} \left(x_{i,k} \cdot \left(\beta_i^{*(l)} \right)_k \right) + \varepsilon_i^{*(l)}. \qquad (10)$$

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，空間的異質性を考慮する．ilr 変換の等長性より，残差 $\mathbf{u}_i \in \mathcal{S}^D$の二乗和は式 (11) 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（Pawlowsky-Glahn et al., 2015），第 2 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．

$$\sum_{i=1}^{n} \|\mathbf{u}_i\|_{\mathrm{A}}^2 = \sum_{i=1}^{n} \|\mathbf{u}_i^*\|^2 = \sum_{i=1}^{n} \sum_{l=1}^{D-1} \left(\mathrm{u}_i^{*(l)} \right)^2. \qquad (11)$$

第 l 番目のパラメータとバンド幅は，式（2, 4）と同様に得ることができる（式 12, 13）：

$$\widehat{\boldsymbol{\beta}}_i^{*(l)} = \left[\mathbf{X}' \mathbf{G}_i(b^{(l)}) \mathbf{X} \right]^{-1} \mathbf{X}' \mathbf{G}_i\left(b^{(l)}\right) \mathbf{y}^{*(l)}, \qquad (12)$$

$$\mathrm{CV} = \sum_{i=1}^{n} \left(y_i^{*(l)} - \sum_{k=1}^{K+1} \left(x_{i,k} \cdot \left(\hat{\beta}_{-i}^{*(l)} \right)_k \right) \right)^2. \qquad (13)$$

以上より，ユークリッド空間における GWR のパラメータ$\widehat{\boldsymbol{\beta}}_i^{*(l)}$と残差$\widehat{\mathbf{u}}^{*(l)} = \mathbf{y}^{*(l)} - \mathbf{X}\widehat{\boldsymbol{\beta}}_i^{*(l)}$が得られる.

最後に, 誤差相関を考慮する. 第 l 番目と第 $\acute{l}$ 番目のモデルの誤差相関は式（14）で表すことができる.

$$\mathrm{E}\left(\varepsilon_i^{*(l)} \varepsilon_j^{*(\acute{l})\prime} \right) = \begin{cases} \sigma_{l,\acute{l}} & (i = j), \\ 0 & (\text{otherwise}), \end{cases} \qquad (14)$$

ここで，$l, \acute{l} \in \{1,2,\dots,(D-1)\}$，$i, j \in \{1,2,\dots,n\}$である．また，$\tilde{\boldsymbol{\varepsilon}}^* = \left(\boldsymbol{\varepsilon}^{*(1)\prime}, \boldsymbol{\varepsilon}^{*(2)\prime}, \dots, \boldsymbol{\varepsilon}^{*(D-1)\prime} \right)'$ とおくと，この $n(D-1) \times 1$の誤差ベクトルの分散共分散行列は式（15）と表すことができる.

$$\boldsymbol{\Omega}^* = \mathrm{E}(\tilde{\boldsymbol{\varepsilon}}^* \tilde{\boldsymbol{\varepsilon}}^{*\prime}) = \boldsymbol{\Sigma}^* \otimes \mathbf{I}_n \qquad (15)$$

ここで，$\boldsymbol{\Sigma}^*$は第 $(l, \acute{l})$ 要素が$\sigma_{l,\acute{l}}$である$(D-1) \times (D-1)$行列であり，$\otimes$はクロネッカー積である．$\boldsymbol{\Omega}^*$は未知であるから，その推定値 $\widehat{\boldsymbol{\Omega}}^*$を式（12, 13）から得られた$\widehat{\boldsymbol{\beta}}_i^{*(l)}$と$\widehat{\mathbf{u}}^{*(l)}$より求める．いま，$\tilde{\mathbf{G}}_i = \mathbf{I}_{D-1} \otimes \mathbf{G}_i$，$\tilde{\mathbf{y}}^* = \left(\mathbf{y}^{*(1)\prime}, \mathbf{y}^{*(2)\prime}, \dots, \mathbf{y}^{*(D-1)\prime} \right)'$，$\tilde{\mathbf{X}} = \mathbf{I}_{D-1} \otimes \mathbf{X}$，$\tilde{\boldsymbol{\beta}}_i^* = \left(\boldsymbol{\beta}_i^{*(1)\prime}, \boldsymbol{\beta}_i^{*(2)\prime}, \dots, \boldsymbol{\beta}_i^{*(D-1)\prime} \right)'$とおくと，実行可能な推定量は式（16）で与えられる：

$$\tilde{\boldsymbol{\beta}}_i^* = \left[\tilde{\mathbf{X}}'\tilde{\mathbf{G}}^{\frac{1}{2}}\hat{\boldsymbol{\Omega}}^{*-1}\tilde{\mathbf{G}}^{\frac{1}{2}}\tilde{\mathbf{X}}\right]^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{G}}^{1/2}\hat{\boldsymbol{\Omega}}^{*-1}\tilde{\mathbf{G}}^{1/2}\tilde{\mathbf{y}}^*. \quad (16)$$

得られた推定値に式（8）の逆変換を行うことで，説明変数の偏弾力性（式 17）と予測確率 $\hat{\boldsymbol{y}}_i$ を導出することができる（Egozcue et al., 2011, Morais et al., 2018）.

$$\text{Semi elasticity}_{i,m} = \left(\ln\left((\beta_i)_{k,m}\right) - \sum_{m=1}^{D}\left(y_{i,m}\ln\left((\beta_i)_{k,m}\right)\right)\right)y_{i,m}. \quad (17)$$

４．所得分析への応用

4.1. 概要

本節では GWSURcoda 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（組成データ）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．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．

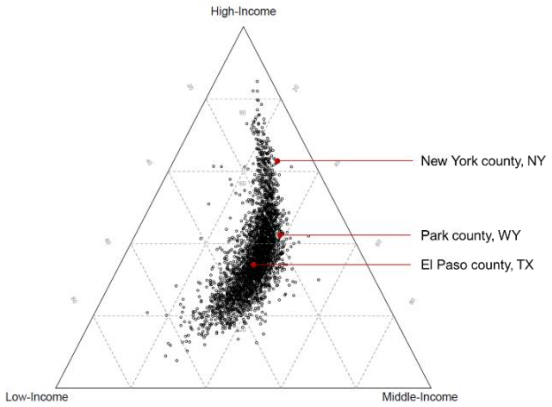

図 1: 所得階層別世帯数割合の三角図

4.2. 結果

推定された各偏弾力性を図 2 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変化率を示す．組成データ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+0.531%，中所得比率は–0.262%，低所得比率は–0.269%となった．同様に，El Paso 郡では，高所得比率は+0.539%，中所得比率は–0.077%, 低所得比率は–0.462%であり, Park 郡では，高所得比率は+0.288%，中所得比率は–0.154%, 低所得比率は–0.134%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 2 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 3 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の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．

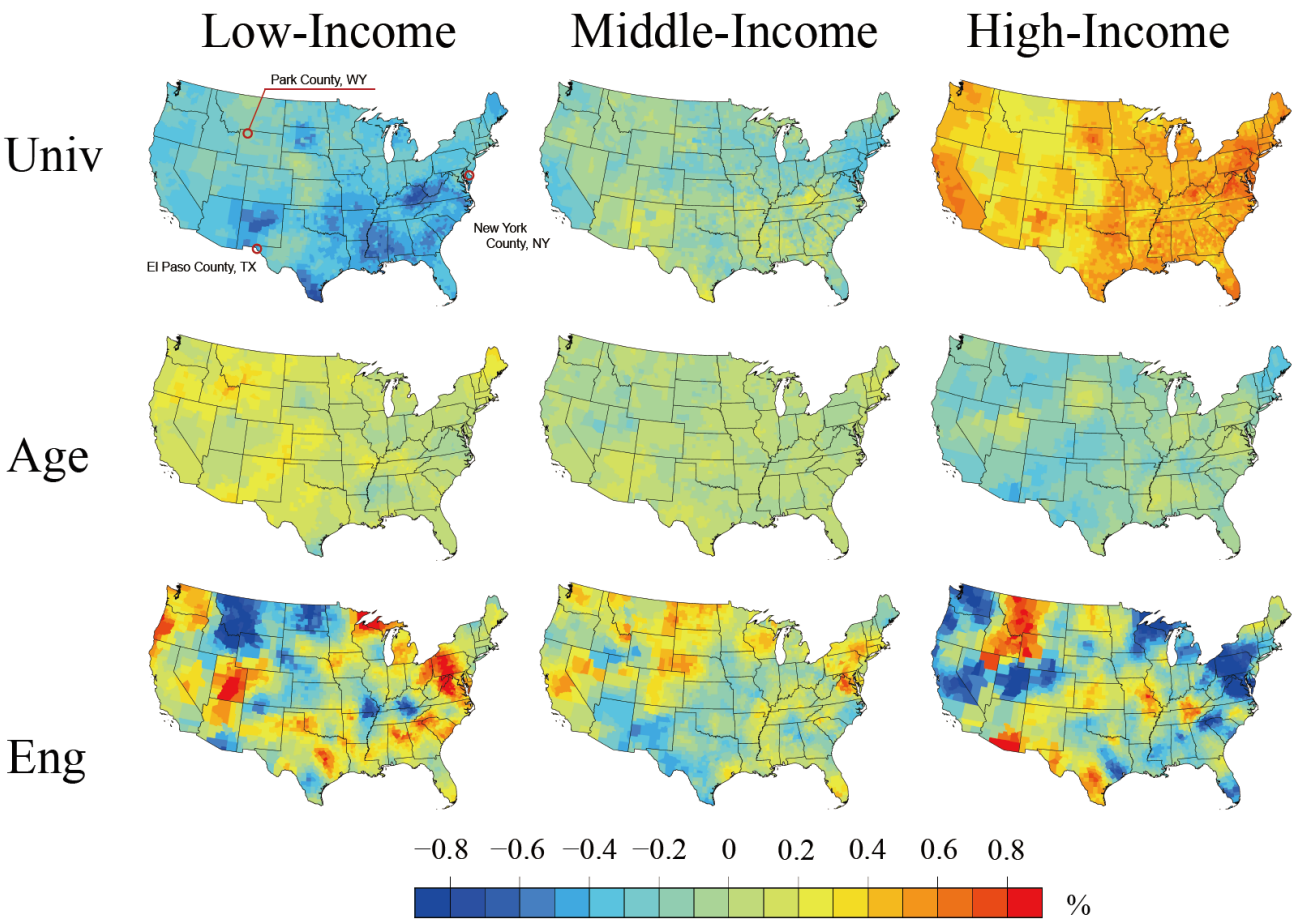

図 2: 偏弾力性の空間分布

図 3: 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）

謝辞

本研究は，データサイエンス共同利用基盤施設の公募型共同研究 DS-ROIS-JOINT（課題番号：005RP2021），ならびに JSPS 科研費（課題番号：17K18554，18H03628，21K13153）の助成を受けた．

参考文献

堤田成政・吉田崇紘・村上大輔・中谷友樹（2021）地理的加重法の研究動向と今後の展望．「GIS－理論と応用」，**29**（1），11–21.

吉田崇紘・村上大輔・瀬谷創・堤田成政・中谷友樹・堤盛人（2020）組成データのための地理的加重回帰モデル．「地理情報システム学会講演論文集」（CD-ROM），**29**，C-21-1-4.

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

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.

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.

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. (2021) Scalable GWR: A linear-time algorithm for large-scale geographically weighted regression with polynomial kernels. *Annals of the American Association of Geographers*, **111** (2), 459–480.

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.

Zellner, A. (1962) An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. Journal of the American Statistical Association, **57**, 348–368.