組成データに基づく地域分類手法の提案

石川理人*・井上　亮**

Regional Classification Methods Based on Compositional Data

Masato Ishikawa*, Ryo Inoue**

Abstract
We propose three methods for unsupervised classification of regions based on compositional data, which is defined as vectors with strictly positive components whose sum is constant. Our methods are extension of the contiguity-constrained hierarchical clustering, which has been used for multidimensional data, to compositional data analysis. We apply our methods to the vote share of political party in 2019 Japanese House of Councillors election in the Kanto region. The proposed three methods captured different aspects of regional characteristics and we found that they complement each other.

Keywords: 組成データ（compositional data），地域分類（regional classification），階層クラスタリング（hierarchical clustering），対数比変換（log-ratio transformation）

1. 序論

組成データとは, ある対象を分類したときの各成分の割合を表すデータである. 地理空間情報には組成データの形式をとるものも多く存在する. 例えば，男女比や，産業別就業者数割合，選挙における政党別得票率は, 都道府県や市区町村といった各領域で異なる値をとる組成データである. 複数の成分で構成される組成データでは, 各領域にスカラーが対応するデータと比較して空間分布の特徴が捉えにくい. 例えば，スカラーデータでは, 各領域の値をコロプレスマップで表現することによって, データの分布や変化点を一目で把握することが可能であるが, 組成データについて同様の表現をするためには, 各成分に対応する複数のコロプレスマップを用いる必要があり, 組成全体の傾向が把握しづらい. このことから，組成データの空間分布から特徴を抽出し, 分かりやすい形で表現する手法が必要である.

データの空間分布に基づく地域の特徴把握手法の１つとして，地域分類が挙げられる. 市区町村などの各領域に１つの値やベクトルが対応したデータについて, 類似する領域を集約して１つのグループにまとめ, データが変化する境界を検出する. 組成データに基づく地域分類は，さまざまな分野への活用が期待できる. 例えば，政党別得票率データに対して地域分類を行えば, 有権者の投票傾向が変化する境界が検出され, 投票先の決定要因に関する考察ができる. また，就業者や学生といった属性ごとの人口割合データに対して地域分類を行えば, 観光地や研究機関の存在が各属性の割合に与える影響について考察することができる. しかし，これまで組成データに基づく地域分類に関する議論は行われていない.

*　学生会員　東北大学大学院情報科学研究科（Tohoku University）
〒980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-06　E-mail：masato.ishikawa.r4@dc.tohoku.ac.jp
**　正会員　東北大学大学院情報科学研究科（Tohoku University）

地域分類を扱ったこれまでの研究では，クラスタリングを用いた手法が多く提案されてきた．最も単純なアプローチは，一般的なクラスタリング手法をそのまま適用するというもの（例えば，Fovell & Fovell (1993)）であるが，この手法では，空間的に連続しない「飛び地」が多く生成する．空間的に連続しないグループを，分類後にすべて別のグループとして設定し直すという対応が考えられるが，分類数を制御しづらいという問題がある．空間的近接性と類似性の双方を考慮した目的関数を最小化するアプローチも複数提案されており，例えば Oliver & Webster (1989) は，空間距離で重みづけした類似度に基づく非階層クラスタリング手法を提案した．ただし，このようなアプローチでは，空間的近接性の考慮の度合いを示すハイパーパラメータを，分析者が試行錯誤して決定する必要がある．

別のアプローチとして，一般的なクラスタリング手法の処理の過程において，空間的近接性を制約として与えることも考えられる．Guo (2008) や Pedroso et al. (2010) は，階層クラスタリングに，隣接したグループのみを結合対象とする制約を与えることを提案した．この手法では，分析者が試行錯誤をすることなく，飛び地が生じない分類を行うことができる．また，各分類数に対応する結果が，ただ 1 つ定まるという利点もある．

本論文では，隣接制約付きの階層クラスタリングによって，組成データに基づく地域分類を行う手法を提案する．組成の類似性を定義した上で，「重心法」，「完全連結法」，「Ward 法」の 3 種類の階層クラスタリング手法に，隣接したグループのみを結合対象とする制約を与え，分類を行う．なお，「重心法」に隣接制約を与えた検討例は，筆者らの知る限り存在しないが，「完全連結法」は Guo (2008)，Pedroso et al. (2010) において，「Ward 法」は Guo (2009) において，それぞれ隣接制約を与えた地域分類が検討されている．

本論文の構成は以下の通りである．第 2 章では，組成の類似性の定義と，具体的な地域分類の手法について述べる．第 3 章では，実データへの適用によって，提案した各手法による結果を比較し，それぞれの手法の特徴について考察する．第 4 章では，本研究の成果と今後の課題を述べる．

2. 提案手法

提案手法では，定義した組成の類似性をもとに，隣接制約付きの階層クラスタリングによって地域分類を行う．階層クラスタリング手法は「重心法」，「完全連結法」，「Ward 法」の 3 種類とし，それぞれの手法に隣接したグループのみを結合対象とする制約を与えて分類を行う．以下に，組成の類似性の定義，および具体的な地域分類の手法を示す．

2.1 組成の類似性

組成を構成する D 個の成分のうち，$D-1$ 個が与えられると，全成分の総和が定数となる組成データの性質によって，残りの 1 成分は自動的に定まる．よって，組成データは，組成を構成する成分数よりも次元が 1 つ少ない単体空間に属している．このことから，各成分を要素とする D 次元のベクトルに基づく分析処理を行うことは適切ではなく，何らかの方法によって，組成データを単体空間から実空間に射影することが必要になる．

Aitchison (1986) は，組成データを実空間に射影する手法として，CLR 変換（有心対数比変換）を提案した．CLR 変換では，組成を構成する各成分について，全成分の幾何平均との比の対数をとる．

$$CLR(\boldsymbol{x}) = \ln \frac{\boldsymbol{x}}{\sqrt[D]{\prod_{d=1}^{D} x_d}} = \ln \boldsymbol{x} - \frac{\mathbf{1}}{D} \sum_{d=1}^{D} \ln x_d \quad (1)$$

ここで，$\boldsymbol{x}$ は組成の各成分を要素とするベクトル，x_d は各成分，D は成分数である．変換によって得られるベクトルは，各成分の分布が正規分布によく近似できるとともに，一般的な比例尺度数値データと同様の扱いが可能になる．組成データの扱いに関する注意点や分析手法については，太

田・新井（2006）が詳しい.

Aitchison (1992) は, 組成の類似性を表す距離関数が満たすべき性質として, ①異なる組成間の距離は正となること, ②同じ組成間の距離はゼロとなること, ③比較する組成間の順序に依らないこと, ④スケール不変性, ⑤摂動不変性, ⑥順列不変性を挙げ, CLR 変換後のベクトル間のユークリッド距離が, それらの条件すべてを満たすことを示した.

$$dist_A(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||CLR(\boldsymbol{x}_i) - CLR(\boldsymbol{x}_j)||_2 \qquad (2)$$

上記の指標は, アイチソン距離と呼ばれる. 本研究では, アイチソン距離を組成の類似性とし, それに基づく地域分類を行う.

2.2 地域分類手法

以下, データを構成する各要素をデータ点と呼ぶ. 階層クラスタリングでは, 類似するデータ点から順に結合してデータを分類する. データ点を結合するとグループが形成されるが, グループ間の類似性評価には複数の手法が存在する. 本研究では,「重心法」,「完全連結法」,「Ward 法」の 3 種類の手法に対して隣接制約を与え, 地域分類を行う. 重心法では, グループ間の重心間距離で類似性を評価する. 完全連結法では, 比較する 2 つのグループに属するデータ点のうち, 最も類似しないデータ点間の距離で類似性を評価する. Ward 法では, グループ結合後のデータ分散で類似性を評価する. そして, 隣接するグループのうち類似するものから順に結合する.

グループを結合する際の基準の違いによって, 得られる結果の性質も変化する. 重心法では, グループ間の重心間距離の最小値を最大化しているため, グループ間でデータの平均的な性質が大きく異なる. しかし, 同じグループに属するデータ点がすべて類似していることは保証の限りではない. 一方, 完全連結法では, 各グループのデータ分布幅の最大値を最小化しているため, 同じグループに属するデータ点であれば, ある程度の類似度が保証される. 各グループのデータ分散の最大値が最小化される Ward 法と特徴が似ているが, 多少類似性が低くても, 少ないデータ点で構成されるグループが, より多くのデータ点で構成されるグループに取り込まれる可能性があるという点で, Ward 法は完全連結法と異なる. これらのことから, 重心法では, 局所的に周囲と性質が大きく異なる地域が, 周囲と異なるグループとして残りやすい. 反対に, Ward 法では, 小さなグループが残りにくく, 大域的なデータの傾向に基づく分類がなされる. 完全連結法は, 両者の間の特徴を示す.

以上の 3 手法を組成データに対して適用する. ただし, 主に人口割合を対象とする本研究においては, 各グループのデータ平均とグループ全体の組成が一致しない. そこで, 重心法においては, 形成された各グループにおいて, グループ全体の組成を求め, これを再度 CLR 変換して重心とする.

3. 実データへの適用

3.1 対象

2019 年の参議院選挙の政党別得票率を対象として, 提案手法による地域分類を行う. 対象地域は, 関東地方の 1 都 6 県のうち, 島嶼部を除く 346 市区町村とし, 自由民主党, 立憲民主党, 公明党, 日本共産党, 日本維新の会, その他の各得票率からなる 6 成分で構成された組成が, 各市区町村に対応する. そして, それぞれの組成に対して CLR 変換を行った上で, 隣接制約付きの 3 種類の階層クラスタリング手法によって地域分類を行う. 対象地域全体の組成は, 図 1 に示す通りである. 各手法による結果を比較する際の参考として, 以下の 3 つの評価指標を設定する.

指標 A：
隣接するグループ間の重心間距離の最小値. 重心法で最大化される.

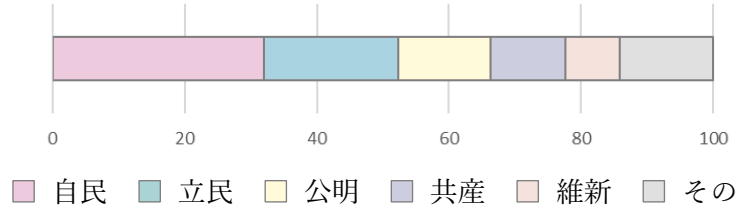


図 1　対象地域全体の組成

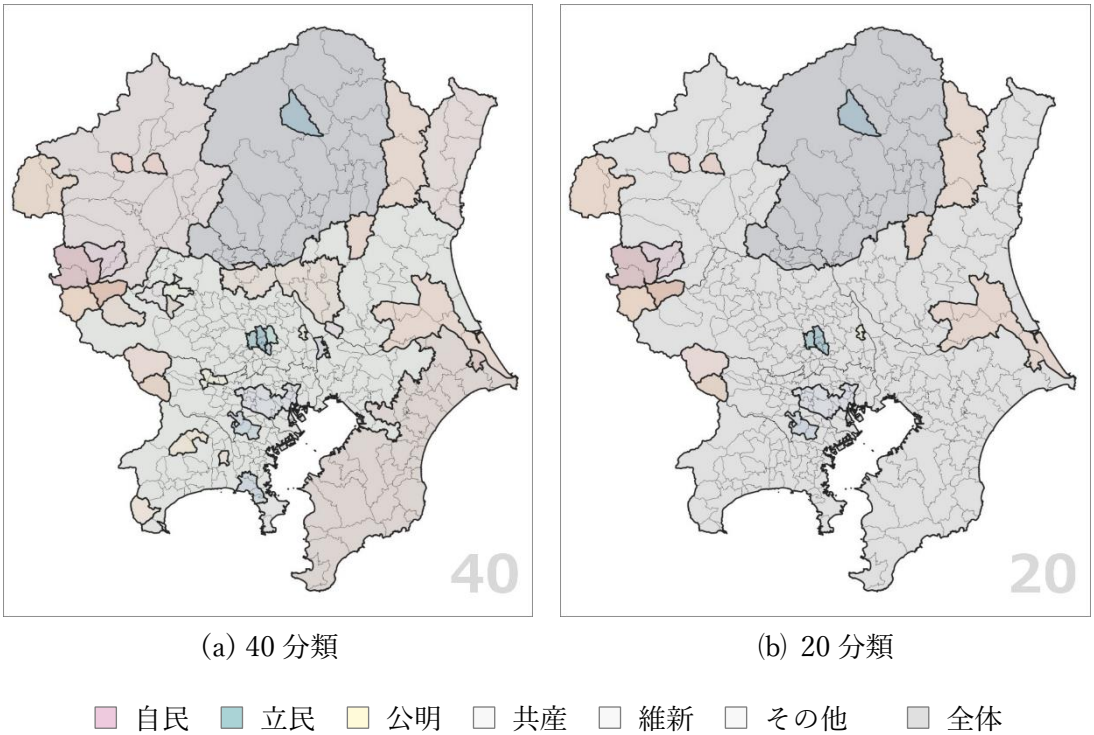
指標 B：
各グループのデータ分布幅の最大値. 完全連結法で最小化される.

指標 C：
各グループのデータの標準偏差の最大値. Ward 法で最小化される. 指標間の次元をそろえるため, 分散ではなく標準偏差とする.

3.2　結果

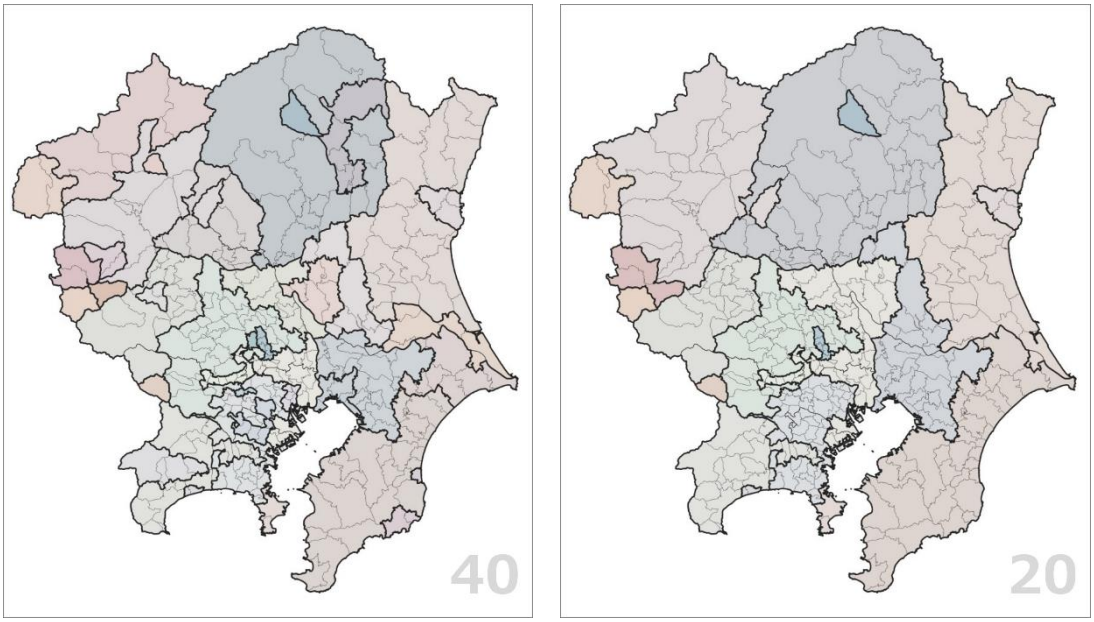
各手法はいずれも, 分類数が領域数と等しい初期状態から, 分類数が 1 となるまでの, 各分類数に対応する結果を出力する. ここでは, 分類数が 40 と 20 の場合の結果を図 2, 図 3, 図 4 に示す. 太線は, 形成されたグループの境界を表す. また, 得票率が大きい, 自由民主党, 立憲民主党, 公明党の 3 成分に対しそれぞれ色を対応させ, 各グループの全体組成を, 3 成分の得票率に応じて混色して表現している. その際, 図 1 に示す対象地域全体の組成を基準とし, 3 成分の得票率が対象地域全体の得票率と等しいグループは灰色で表現される.

各手法による結果から, 得票率の組成が県境で変化しやすい傾向を把握することができる. 参議院選挙は原則として各都道府県が 1 つの選挙区となり, 各選挙区の立候補者に対して投票する選



(a) 40 分類　　(b) 20 分類

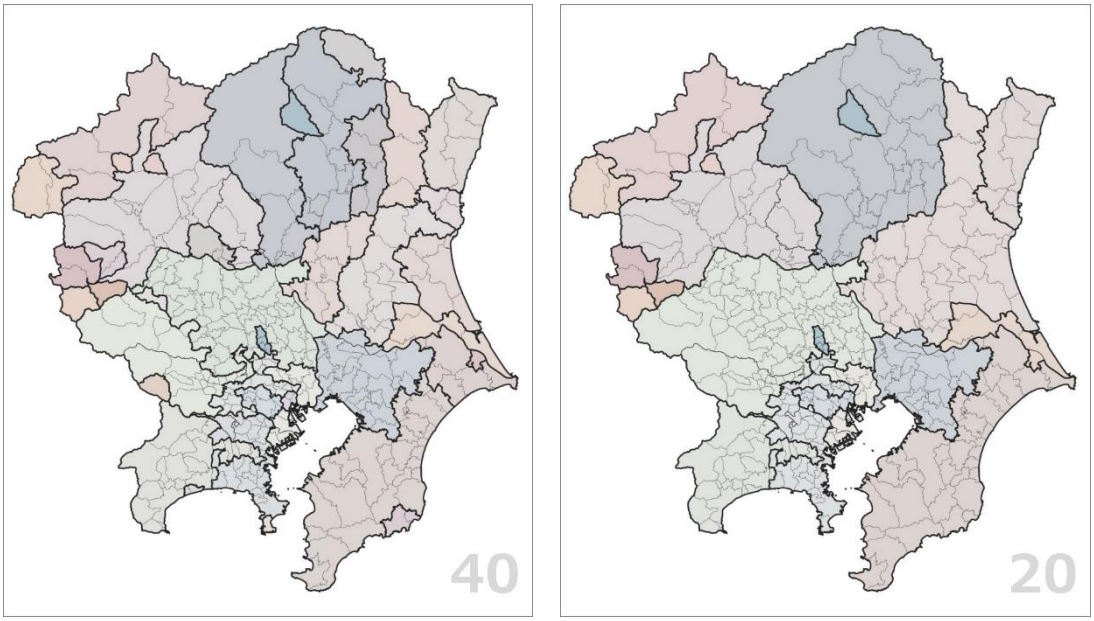
図 2　重心法による結果



(a) 40 分類　　(b) 20 分類

自民　立民　公明　共産　維新　その他　全体

図 3　完全連結法による結果



(a) 40 分類　　(b) 20 分類

自民　立民　公明　共産　維新　その他　全体

図 4　Ward 法による結果

挙区選挙と, 選挙区に関係なく政党別に投票する比例代表選挙が行われる. 今回対象としたデータは後者のみであるが, 県境で組成が変化しやすい傾向が確認された. このことから, 形式上は選挙区が関係しない比例代表選挙でも, 選挙区が各政党の得票率に影響を与えていることが示唆される.

重心法では, 南関東一帯が広大な 1 つのグループに集約されるとともに, 東京都心部やさいたま市, 栃木県全域などが, それぞれ南関東一帯とは異なるグループとして残っている. 局所的に周囲と性質が大きく異なる地域が, 周囲と異なるグループとして残りやすい, この手法の特徴を確認することができる. 一方, 完全連結法や Ward 法では, 各グループの大きさがより均等になっている. 双方の結果は類似している部分もあるが, 20 分類において, 前者では埼玉県周辺がいくつかのグループに分けられているのに対し, 後者では 1 つのグループにまとめられているなど, 相違点も多い. 細部の大きな変化に敏感な完全連結法と, より大域的な傾向が反映されがちな Ward 法の性質の違いが, このような結果の違いをもたらしていると考えられる.

各結果に対応する評価指標 A, B, C を表 1 に示す. 指標 A は値が大きいほど, 指標 B, 指標 C は値が小さいほど, それぞれの観点において適切な分類がなされていると判断できる. 指標 A は重心法で, 指標 B は完全連結法で, 指標 C は Ward 法でそれぞれ最適化され, 実際にそれぞれの指標は対応するそれぞれの手法で最もよい値をとっていることが確認できる. 一方, 指標 A は, 完全連結法で小さい値をとっており, 隣接するグループの全体組成が類似していても, 両者が別々のグループとして分類されている部分があることを意味している. 指標 B は, 重心法で大きい値をとっており, 広い範囲にわたって徐々に組成が変化している領域が集約され, 分布幅の広いグループが形成されていることが推測される. 指標 C は, Ward 法以外の 2 手法で大きな値をとっており, 分類数によって大小が逆転している. このように, 各手法にはそれぞれ利点と欠点が存在し, すべてを同時に最適化する手法は存在しないとともに, 各指標に基づく手法間の優位性は, 分類数や対象とするデータによって変化し得る. よって, 利用する手法を 1 つに限定するのではなく, 地域分類の目的や指標の値に応じて, 適切な手法を選んだり, 各手法による結果を比較して考察したりすることが重要である.

4. 結論

本論文では, 組成データに基づく 3 種類の地域分類手法を提案した. 提案手法は, アイチソン距離により評価した組成の類似性をもとに, 隣接制約付きの 3 種類の階層クラスタリングを行うことにより, 組成が類似する市区町村などの領域を集約し, 組成が変化する境界を検出する. 実データへ適用した結果, 提案手法を用いた分類によって, 組成データの空間的な分布の特徴が容易に把握できるとともに, 各手法にはそれぞれ特有の性質が存在することが確認された.

提案手法には課題も存在する. 人口割合を対象

表 1　各指標

	40			20		
	指標 A	指標 B	指標 C	指標 A	指標 B	指標 C
重心法	**0.457**	1.166	0.340	**0.584**	1.554	0.425
完全連結法	0.142	**0.708**	0.309	0.145	**0.917**	0.489
Ward 法	0.186	1.019	**0.266**	0.320	1.233	**0.312**

とした分析を行う場合, 各領域の組成は人口規模の多寡の影響を受ける. 人口が多い領域では, 組成を構成する各成分の割合は安定するが, 人口が少ない領域では, 各成分の割合は不安定になる. 特に, ゼロに近い成分が存在する場合, その成分に属する人口が数人増減しただけで, CLR 変換後のベクトルは大きく変化する. このような組成の「ぶれ」は, 分類の結果に大きく影響する可能性がある. よって, 人口が少ない領域における組成が分類に与える影響を小さくする補正を, 元データに対して, もしくは結合処理の段階で行う必要があるが, その具体的な手法については検討できていない. また, 階層クラスタリングに基づく提案手法は, データの階層構造が把握できる反面, それぞれの分類数に対して最適な結果が得られない. よって, 非階層クラスタリングに隣接制約を与えた手法などを今後検討し, 階層制約の有無が結果にもたらす影響について考察する必要がある.

謝辞

本研究は, JSPS 科研費 21H01447 の助成を受けた.

参考文献

太田亨・新井宏嘉 (2006) 組成データ解析の問題点とその解決方法.「地質学雑誌」, **112** (3), 173-187.

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Aitchison, J. (1992) On criteria for measures of compositional differences. *Mathematical Geology* **24**, 365–379.

Fovell, R. G. and Fovell, M. -Y. C. (1993) Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate*, **6** (11), 2103－2135.

Guo, D. (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, **22** (7), 801-823.

Guo, D. (2009) Greedy optimization for contiguity constrained hierarchical clustering. *2009 IEEE International Conference on Data Mining Workshops*, 591-596.

Oliver, M. A. and Webster, R. (1989) A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, **21**, 15–35.

Pedroso, M., Taylor, J., Tisseyre, B., Charnomordic, B. and Guillaume, S. (2010) A segmentation algorithm for the delineation of agricultural management zones. *Computers and Electronics in Agriculture*, **70** (1), 199-208.