

令和 5 年 6 月 15 日現在

機関番号：12102

研究種目：挑戦的研究（萌芽）

研究期間：2017～2022

課題番号：17K18554

研究課題名（和文）組成データ解析の新たな展開

研究課題名（英文）New developments in compositional data analysis

研究代表者

堤 盛人（Tsutsumi, Morito）

筑波大学・システム情報系・教授

研究者番号：70292886

交付決定額（研究期間全体）：（直接経費） 4,800,000円

研究成果の概要（和文）：割合などのように値が非負で和が一定なデータである「組成データ」の分析の際に関し、約40年前から、値の総和が一定であるという定数和制約を考慮した「組成データ解析（CoDA）」が発展してきているが、その適用の大半は自然科学データである。本研究では、組成データ解析の社会経済データへの応用を主眼に、既存の統計分析手法を俯瞰・体系化し、その新たな展開の可能性を探ることを目的として、社会経済分析におけるCoDAの意義の明確化、空間計量経済学におけるCoDAの可能性の検討、時間軸を考慮した組成データを用いた理論的拡張の可能性の検討、そして、組成データ並びにCoDAの結果の新たな視覚的表現手法の検討を行った。

研究成果の学術的意義や社会的意義

現在ではCoDAに関わる研究者集団も組織されている。（<http://www.compositionaldata.com/>）しかしながら、これまでCoDAの研究において取り扱われているのは自然科学データを用いた研究が大半であり、社会経済分析の分野でその重要性・有用性が十分認識されるに至っていなかった。本研究においては社会経済分析に着目していくつかの研究成果を上げている。これにより、分野を超えた知見の共有が可能となり、これまでのCoDAの枠組みを変える展開、新たな学問的発展に寄与することが期待出来る。

研究成果の概要（英文）：Compositional data analysis (CoDA), which takes into account the constant sum constraint that the sum of values is constant, has been developed for the analysis of compositional data, which are data with non-negative values and constant sums, such as proportions, for about 40 years, but the majority of its applications are to natural science data. In this study, with the main focus on the application of compositional data analysis to socio-economic data, we aim to overview and systematise existing statistical analysis methods and explore the potential for their new development: clarifying the significance of CoDA in socio-economic analysis, examining the potential of CoDA in spatial econometrics, and considering the potential of compositional data considering a time axis. The study also examined the possibility of theoretical extensions using compositional data that take into account the time axis, and new visual representation methods for compositional data and CoDA results.

研究分野：土木計画学

キーワード：組成データ CoDA 社会経済データ 地理空間データ

1. 研究開始当初の背景

割合などのように値が非負で和が一定となるようなデータは「組成データ (Compositional Data)」と呼ばれている。その名称自体は一般的には知られていないものの、至る所で目にするデータの種類である。普段あまり意識はされないが、統計学的には、疑似相関の問題から組成データの分析の際には値の総和が一定であるという定数和制約を考慮する必要があり、地質学を中核にこれを考慮した「組成データ解析 (Compositional Data Analysis: CoDA)」が発展している。定数和制約により変数の自由度が一つ下がることから、データの定義域が単体空間上に限定される。CoDAはこの単体空間上の統計解析手法を提供している。

しかしながら、この分野のバイブルである Aitchison (1986) *The Statistical Analysis for Compositional Data*, Chapman and Hall から既に 30 年が経過しているにも関わらず、未だ CoDA の研究において取り扱われているのは自然科学データが大半で、地理学や社会科学ではその重要性・有用性がほとんど認識されておらず、地理的な社会経済データを用いた実証研究は、ごくわずかな限られた研究者による例外を除いて皆無に近い。組成データがありふれたデータであることと、統計ソフトなどの上では上述の統計学的な問題を無視したままでも通常の統計解析が適用可能なことから、今なお、組成データの特性が理解されないまま問題が見過ごされることが多い。

さらに、地理空間データの特性 (空間的自己相関・空間的異質性) を考慮した空間統計学的手法を用いた CoDA の研究は、地質学に近い鉱山学を起源のひとつにもつ地球統計学 (kriging など) に基づくものがほとんどであり、特に人口や土地利用などの地理空間データの特性を扱う空間計量経済学との融合には大きな研究余地が残されている。

2. 研究の目的

そこで本研究では、地理的な社会経済データへの応用を主眼に、空間統計学的手法に基づいた CoDA の新たな統計解析手法の開発とその実用化を主な目的とする。地理学においては地域特性を示す変数として五歳階級別人口構成比や産業別人口割合などの組成データを扱うことも多いため、組成データの特性とそれを見過ごすことによるデータ分析上の影響を整理することは、地理空間データの解析手法の体系化・高度化という地理学分野における学術的課題への貢献も期待できる。そして、CoDA の解析手法を分野間の隔たりが存在した空間統計学の知見を活用して高度化することで、以下に挙げるような学術的発展が期待できる。

(1) 社会経済分析における CoDA の意義の明確化

地域における産業の相対的な集積度を簡便に計測可能な地域特化係数 (Location Quotient: LQ) のように、比率に基づいた指標は多い。特に、本研究チームの関心領域である地域科学において、LQ のほか Ellison-Glaeser 係数などの集積性を示す指標には比率を用いたものは多い。ここで、それらの指標では、組成データ (定数和制約の掛かったベクトル) の一変量を抜き出して評価したものであると見ることが出来る。つまり、CoDA の観点からいえば、多変量として定数和制約を考慮してベクトルを評価すべきであるものが、恣意的にひとつの要素だけで評価しているものといえる。CoDA の知見を援用してこの改善を図る余地は大きく、それにより、CoDA の意義および有用性を示す材料を提示できる可能性がある。

(2) 空間計量経済学における CoDA の可能性の検討

回帰モデルにおいて、集計型のロジットモデルや線形確率モデルなどの被説明変数に比率をとるモデルは、CoDA の分野では組成モデルと呼ばれ、オッズの対数をとるだけでなく様々な対数比変換法を適用したモデルが整理されている。

組成データを扱っているにも関わらず、長年の慣習から CoDA の知見を活用しない研究もいまだ多くあり、そうした分析は結果の解釈に重大な欠陥が生じることも指摘されている。空間計量経済学手法を用いた一流の論文においても、そうした状況が残っている。このように地理空間データの特性を考慮した CoDA は限られており、本研究で取り組む空間統計学、特に空間計量経済学の知見を活かした新たな方法論の開発は CoDA 研究の一分野を切り拓く可能性がある。

(3) 時間軸を考慮した組成データを用いた理論的拡張の可能性の検討

地理的な社会経済データは、地質学等で用いるデータと比べて、クロス分析に用いるような二次元配置データ、OD 表のようなフローデータ、階層構造をもつ交通機関分担率データ、選挙における政党別得票率データなど、多様かつ多岐にわたる。

これまでの地質学等での CoDA では、主として時間不変な場合 (何故なら、鉱物における化学組成はほぼ時間不変といえる) を対象としており、ごく最近になって時間軸の導入が試みられている段階であり、時間可変なデータについてはいままさに挑戦的な課題と認識されている。さ

らに、社会経済データでは、データの次元数が時間可変な場合(例えば、政党別の得票率データ。政党は分離・併合・創設などが発生する)もある。このように地理空間データに着目することで CoDA の理論的拡張・実用化にも大きな貢献が期待できる。

(4) 組成データ並びに CoDA の結果の新たな視覚的表現手法の検討

組成データは3次元以下であれば、いわゆる三角図によって容易に視覚化可能であるが、4次元以上になると何らかの工夫が必要となり、適切な視覚化手法は見当たらない。クラスター分析等についても、本来の CoDA の特性を考慮した視覚化手法は見当たらない。社会経済データでは、産業分類等のように次元の多い組成データを扱う機会も多く、それらの特徴をわかりやすく表現する視覚的表現手法への期待は大きい。

3. 研究の方法

(1) 社会経済分析における CoDA の意義の明確化

地域特化係数 Location Quotient (LQ) の改良を行う。地域における産業の相対的な集積度を簡便に計測可能な LQ は、地域科学分野において幅広く利用されてきた。LQ は、経済活動の空間分布の分析や地域経済の強みである基盤産業の把握を行う代表的な係数のひとつである(e.g., 黒田ほか, 2008)。地域 i における産業 p の地域特化係数 $LQ_{i,p}$ は式(1)で定義される:

$$LQ_{i,p} = x_{i,p}/g_p \quad (1)$$

ここで、 $x_{i,p}$ は地域 i における産業 p の従業者数シェア、 g_p は全国における産業 p の従業者数シェアである。 $LQ_{i,p} = 1$ であれば、地域 i における産業 p のシェアは全国レベルと同じであることを示す。 $LQ_{i,p} > 1$ であれば、全国平均を超えて地域 i に産業 p が集中していることを示し基盤産業とみなされる。一方、 $LQ_{i,p} < 1$ であれば、非基盤産業とみなされる。

これまで、LQ には様々な改良版が提案されている。式(1)の計算に必要なデータは従業者数のみのため、比較的データが集めやすく計算も容易で解釈もわかりやすいといった利点がある。反面、基盤産業と非基盤産業の区別に用いる閾値の設定に批判があり、この点の改良が長年行われている。例えば、基準化した LQ を提案し、正規性の検定によって区別する方法や、LQ を対数化した上で基準化を行い、bootstrap 法を用いて統計的検定を行う評価方法が提案されている。この他に、LQ は経済波及効果の分析に用いられる産業連関表の交易係数の推定手法としても広く用いられており、推定精度の向上を目的として LQ をベースとした開発が近年注目されている。

本研究ではさらなる LQ の改良に向けて、新たな課題を指摘する。式(1)の分子、分母のシェアは、組成データ(定数和制約の掛かったベクトル)の一変量を抜き出したものである。産業 p のシェアの計算に本来含まれている産業 p 以外のシェアの影響を捨象しており、全産業のシェア率を考慮の上で産業 p に着目する指標になってはいない。そこで本研究では、地質学や鉱山学で用いられてきた組成データに対する解析手法 CoDA を援用し、全産業のシェア率を考慮した新たな地域特化係数を提案する。CoDA を援用することにより、定数和制約に起因する問題に対応し、特定の産業だけでなく産業構成全体の情報も取り入れた新たな地域特化係数を開発することを試みる。

(2) 空間計量経済学における CODA の可能性の検討

組成データの回帰モデルでは、被説明変数も多変量データとなる。空間計量経済学のモデルでは、地点 i の近傍を空間重み行列 W によって定義し、被説明変数 y の近傍の情報 Wy を説明変数として取り込むことで被説明変数の空間的自己相関を考慮する空間ラグモデルがある。変数が一変量の場合は、その変量の空間的自己相関であるが、多変量の場合、変量間の空間的相関「空間的相互相関」を考慮する必要がある。正の空間的相互相関は、具体的に、メッシュ単位の土地利用組成データの場合では「メッシュ内の水域の割合が高いメッシュの近傍メッシュでは、田の割合が高い」ということになる。組成データの場合、これに加えて定数和制約の考慮も必要となる。そこで本研究では、計量経済学における Seemingly Unrelated Regression により多方程式間の誤差相関を考慮した上で、組成データ解析(対数比変換)と空間計量経済学(空間重み行列)を導入したモデルを検討し、これを拡張した新たなモデルを開発する。

さらに、地理的加重回帰(Geographically Weighted Regression: GWR)モデルなどに代表される空間異質性を考慮可能なモデルを、組成データに適用可能となるような高度化の方策を検討する。

(3) 時間軸を考慮した組成データを用いた理論的拡張の可能性の検討

米国大統領選挙、および日本の参議院選挙の複数時点の党別得票率データを用いた地域的な政治的傾向の抽出を、CoDA と時系列データを関数として表現し扱う関数データ解析(Functional Data Analysis)を組み合わせることで定数和制約と時間軸を同時に考慮したクラスタ分析手法を行い、その理論的拡張可能性を検討する。

(4) 組成データ並びに CoDA の結果の新たな視覚的表現手法の検討

CoDA 研究の中心的研究機関であるジローナ大学(スペイン)が、三次元三角図(三角錐の 4

つある頂点を最大値（100 や 1）とし内部に対応する組成をプロットした図）を実装した CoDa Pack (<https://www.compositionaldata.com/codapack.php>) という 4 次元の組成データを視覚化するツールを公開していることを、研究分担者である吉田が同大のワークショップに参加した際に知ったため、この有用性を確認することとする。

4. 研究成果

(1) 社会経済分析における CoDA の意義の明確化

CoDA に基づく演算子を応用し、特定の産業だけでなく産業構成全体を考慮に入れた新たな地域特化係数を提案した。提案した係数を日本の産業分類データに応用し、その地域特化の程度の解釈を検討した。これらの成果を、International Conference on Econometrics and Statistics と地理情報システム学会において発表し、国内外の専門家との意見交換を行った。

成果 では、CoDA の知見を援用することで全産業の構成を考慮した地域特化係数の整理を行えることを明らかとした。成果 では、地域特化係数と提案した係数を軸とする散布図に両係数の閾値（-1, 0, 1）で区切られた領域を設定し、各領域に解釈の意味付けを行うことで、自地域における立地特性を把握可能であることを明らかとした。

Takahiro Yoshida, Daisuke Murakami, Hajime Seya (2022) Location powered quotient: A compositional data analysis-based approach. The 5th International Conference on Econometrics and Statistics (Online and Kyoto, Japan; June 4–6, 2022), ID: E0763.

吉田崇紘・村上大輔・瀬谷創 (2022) 「Location powered quotient: A compositional data analysis-based industrial concentration measure」. 『地理情報システム学会講演論文集』(CD-ROM), Vol.31, 講演番号: B-2-5. 地理情報システム学会第 31 回研究発表大会, 沖縄産業支援センター/オンライン, 2022 年 10 月.

(2) 空間計量経済学における CoDA の可能性の検討

空間的自己相関と空間的相互相関に加え、誤差相関を対処するモデルと空間的異質性を考慮するモデルを、それぞれ検討した。特に後者については、地理的加重回帰モデルを組成データ用に拡張し、米国における所得階級別人口の組成データへの適用を行い、定数和制約を伴った形で得られるパラメータ推定値、予測値の解釈上の有用性を示すことができた。

成果 では、組成データであることから被説明変数が複数であること、また、それにより空間ラグ項も複数になることから、回帰係数を求めるためには連立方程式を解く必要があることを整理し、同モデルでは定数和制約、空間相関、誤差相関を同時に対処する必要性を明らかとした。この成果は、North American Meetings of the Regional Science Association International で発表した。成果 では、CoDA における演算子を用いることで、GWR も単体空間上において定式化できることを明らかとした。成果 では、対数比変換をした上で得られた回帰係数を逆対数比変換して得られた逆変換済回帰係数が組成データとなることを明らかとした。また、その係数が地域ごとに得られることから、マップ化することで空間的な解釈が可能であることを明らかとした。成果 では、と同様に、誤差相関の対処として連立方程式モデルが有効であることを明らかとした。から の成果は、地理情報システム学会と International Conference on Geographic Information Science に発表した。さらに組成データ解析の専門家らとも密に議論するため、組成データ解析の応用例が多い地質データに開発したモデルを適用し、International Association for Mathematical Geosciences での口頭発表が採択されている（発表は 2023 年 8 月に予定）。

Takahiro Yoshida (2019) Spatial seemingly unrelated regression model combined with compositional data analysis approach. The 66th Annual North American Meetings of the Regional Science Association International (Pittsburgh, USA; November 13–16), ID: P160375.

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), Vol.29, 講演番号: C24-1-4. 地理情報システム学会第 29 回研究発表大会, オンライン, 2020 年 10 月.

Takahiro Yoshida, Daisuke Murakami, Hajime Seya, Narumasa Tsutsumida, Tomoki Nakaya (2021) Geographically weighted regression for compositional data: An application to the U.S. household income compositions. Proceedings of the 11th International Conference on Geographic Information Science (Online; September 27–30, 2021. Originally planned: Poznan, Poland; September 15–18, 2020), paper 55. DOI: 10.25436/E2G599.

吉田崇紘・村上大輔・瀬谷創・堤田成政・中谷友樹 (2021) 「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」. 『地理情報システム学会講演論文集』(CD-ROM), Vol.30, 講演番号: C30-1-5. 地理情報システム学会第 30 回研究発表大会, オンライン, 2021 年 10 月.

(3) 時間軸を考慮した組成データを用いた理論的拡張の可能性の検討

CoDA と関数データ解析を組み合わせることで定数和制約と時間軸を考慮したクラスタ分析手法を構築し、米国大統領選挙における選挙得票率データに適用した。郡単位かつ複数時点(2000~2016年までの5時点)のクラスタ分析結果は、人種や貧困率などの社会経済のほか、国境周辺や東部・南部の差異などの地域特性が反映された分類を行うことができた。組成データの時間的な安定性が分類に影響することから、優占党が比較的安定した州単位と、時点によって比較的ばらつきのある郡単位を比較した分析を行い、ポアソン回帰におけるオフセット項のように、組成だけでなく総数も含めて分析することが重要であるとの示唆を得た。この成果は、Spatial Econometrics Association で発表した。

Morito Tsutsumi, Noriyuki Nukaga, Takahiro Yoshida (2021) County-level Geographic Characteristics of the U.S. Presidential Election Outcome: 2000–2016. XV World Conference of the Spatial Econometrics Association (Online; May 26–28, 2021. Originally planned: Tokyo, Japan; May 26–28, 2021) ID: 7-B-04.

(4) 組成データ並びに CoDA の結果の新たな視覚的表現手法の検討

CoDa Pack の有用性を、選挙得票率データを用いて考察した。一方で、平面上では4次元以上の表現が難しいことから、多次元を集約化する方法として、次元の類似性からクラスタを構成する Q-mode クラスタリングにより次元数を最大で4とする方法の有用性についても考察を行った。Q-mode クラスタリングにより、政党の離合集散や諸派の影響について、投票行動から解釈のしやすい分類を行えることを確認した。

また、CoDA の結果の視覚的表現として、Aitchison 距離を用いた地理的な表現の有用性についても確認した。

Yoshida Takahiro, Er-rbib Rim, Tsutsumi Morito (2019) Which Country Epitomizes the World? A Study from the Perspective of Demographic Composition, SUSTAINABILITY/11(22), 2019-11.

5. 主な発表論文等

〔雑誌論文〕 計6件（うち査読付論文 4件／うち国際共著 0件／うちオープンアクセス 0件）

1. 著者名 Yoshida, Takahiro;Murakami, Daisuke;Seya, Hajime;Tsutsumida, Narumasa;Nakaya, Tomoki	4. 巻 -
2. 論文標題 Geographically weighted regression for compositional data: An application to the U.S. household income compositions	5. 発行年 2021年
3. 雑誌名 Proceedings of the 11th International Conference on Geographic Information Science	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 江端 杏奈, 吉田 崇紘, 為季 和樹, 瀬谷 創, 堤 盛人	4. 巻 29 (1)
2. 論文標題 ふるさと納税の探索的空間データ分析	5. 発行年 2021年
3. 雑誌名 GIS-理論と応用	6. 最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 吉田崇紘・村上大輔・瀬谷創・堤田成政・中谷友樹・堤盛人	4. 巻 29
2. 論文標題 組成データのための地理的加重回帰モデル	5. 発行年 2020年
3. 雑誌名 地理情報システム学会講演論文集（CD-ROM）	6. 最初と最後の頁 C24-1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Yoshida Takahiro, Er-rbib Rim, Tsutsumi Morito	4. 巻 11
2. 論文標題 Which Country Epitomizes the World? A Study from the Perspective of Demographic Composition	5. 発行年 2019年
3. 雑誌名 Sustainability	6. 最初と最後の頁 6404 ~ 6404
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/su11226404	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Takahiro Yoshida and Morito Tsutsumi	4. 巻 11(1)
2. 論文標題 On the effects of spatial relationships in spatial compositional multivariate models	5. 発行年 2018年
3. 雑誌名 Letters in Spatial and Resource Sciences	6. 最初と最後の頁 57,70
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s12076-017-0199-5	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 吉田崇紘・堤盛人	4. 巻 25(2)
2. 論文標題 人口構成比の観点からみた将来の日本の縮図：組成データ解析の適用	5. 発行年 2017年
3. 雑誌名 GIS - 理論と応用	6. 最初と最後の頁 23,33
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

[学会発表] 計8件 (うち招待講演 0件 / うち国際学会 6件)

1. 発表者名 Takahiro Yoshida, Hajime Seya
2. 発表標題 On spatial prediction of apartment rent using machine learning approaches
3. 学会等名 XV World Conference of the Spatial Econometrics Association (国際学会)
4. 発表年 2021年

1. 発表者名 MoritoTsutsumi, Noriyuki Nukaga, Takahiro Yoshida
2. 発表標題 County-level Geographic Characteristics of the U.S. Presidential Election Outcome
3. 学会等名 XV World Conference of the Spatial Econometrics Association (国際学会)
4. 発表年 2021年

1. 発表者名 吉田崇紘・村上大輔・瀬谷創・堤田成政・中谷友樹
2. 発表標題 定数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
3. 学会等名 地理情報システム学会第30回研究発表大会, オンライン
4. 発表年 2021年

1. 発表者名 Anna Ebata and Morito Tsutsumi
2. 発表標題 A Spatial Analysis of Industrial Concentration in Japan Using Compositional Data: An Application of Geographically Weighted Spatial Statistics
3. 学会等名 XII World Conference of the Spatial Econometrics Association (SEA) (国際学会)
4. 発表年 2018年

1. 発表者名 Takahiro Yoshida and Morito Tsutsumi
2. 発表標題 Compositional multivariate conditionally autoregressive (CMCAR) model with spatial cross-correlation for discrete space
3. 学会等名 Spatial Statistics 2017 (国際学会)
4. 発表年 2017年

1. 発表者名 Takahiro Yoshida, Rim Er-Rbib, and Morito Tsutsumi
2. 発表標題 The epitome of the future world from the perspective of demographic composition
3. 学会等名 The 7th International Workshop on Compositional Data Analysis (国際学会)
4. 発表年 2017年

1. 発表者名 Takahiro Yoshida and Morito Tsutsumi
2. 発表標題 Spatial regression model for compositional data in discrete space with spatial auto-correlation and spatial cross-correlation
3. 学会等名 The 7th International Workshop on Compositional Data Analysis (国際学会)
4. 発表年 2017年

1. 発表者名 山口真理沙・吉田崇紘・堤盛人
2. 発表標題 過去日本の縮図の探索：五歳階級別人口構成比の観点から
3. 学会等名 土木学会第55回土木計画学研究発表会（春大会）
4. 発表年 2017年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究 分担者	吉田 崇紘 (Yoshida Takahiro) (60826767)	東京大学・空間情報科学研究センター・助教 (12601)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------