研究ノート

階層的クラスター分析結果の新たな解釈手法：「変数別寄与率算定法（LIV 法）」の提案
～家計調査クラスター分析結果への適用～

木村　敦 | Kimura Atsushi

㈱統計センター 理事・CIO

■ NTT にて ICT 関連開発に長年携わり、㈱ NTT ファシリティーズ総合研究所 取締役情報技術本部長を経て、2019 年 4 月から現職。1988 年 3 月名古屋大学大学院理学研究科博士課程（前期）修了、修士（理学）、専門統計調査士。

1. はじめに

木村・高部 [1] では、家計調査データを用いたクラスター分析によって、日本における日常的食料嗜好の地域性について明らかにするとともに（図 3、4）、213 食料品目のうち地域性に大きく寄与すると思われる品目について数値的な簡易解釈結果（2 値ヒートマップ）を示した。

本稿では、木村・高部 [1] で報告された食料嗜好地域性の発現に寄与する食料品目について、より詳細に定量分析を行うための手法の提案を行うとともに、提案手法を木村･高部 [1] のクラスター分析結果の解釈に適用する。

階層的クラスター分析結果の解釈は主観的に行われることが多い。クラスター凝集に対して、具体的にどの変数が強く寄与したのかを把握するための客観的な手法として決定木などを用いる方法もあるが、両者のアルゴリズムが本質的に異なることから必ずしも納得性の高い解釈結果が得られるとは限らない。ヒートマップも元データの全体的傾向を把握するために用いられるが、クラスターの凝集と観測変数の関係を直接的に示すものではないことに加えて元データの変数の数が多くなると直感的に理解することが難しくなるのが難点である（図 1）[注1)]。

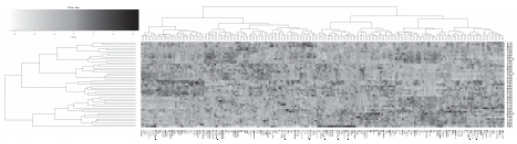

図 1　木村・高部 [1] の結果を HeatMap にしたもの

本稿で新たに提案する「変数別寄与率算定法」の基本的な考え方は、階層的クラスター分析として良く用いられる ward 法（Ward [2]）の凝集アルゴリズムに着目して、クラスターの凝集・分割に対して各変数が寄与した度合いを「変数別情報損失寄与率（LIV: Loss of Information rate by Variable）」と名付けた数値を用いて定量的に評価するものである。この「変数別寄与率算定法」を木村・高部 [1] の分析結果に適用したところ、凝集の初期工程では双方の地域の特産品などの LIV 値が高くなり、凝集が進むにつれて地域個別の特産品などの効果が平均化されクラスター内要素全体の共通的な傾向が浮き出てくることを定量的に明らかにすることができた。

なお本論文における見解は筆者個人のものであり、所属する組織を代表するものではない。本文

注１）　図１及び図５～図 10 のカラー版と、「標準化データ A」から各パレート図を作成するエクセルシートを、(公財)統計情報研究開発センター HP（https://www.sinfonica.or.jp/）内の［刊行物］＞［エストレーラ］＞［参考］に掲載した。

章の誤りはすべて筆者の責に帰する。

2. ward 法の概要と変数別分離寄与の考え方

⑴ ward 法の凝集アルゴリズム

ward 法におけるクラスター凝集のアルゴリズムについて簡単に述べておく。凝集過程における i 番目のクラスター i について、要素数を Ni、クラスター i の j 番目の要素ベクトルを $\boldsymbol{x_{ij}}$、クラスター i の重心ベクトルを $\boldsymbol{c_i}$ とする。各要素における観測データ変数の数を m とすれば、$\boldsymbol{x_{ij}}$ も $\boldsymbol{c_i}$ もともに m 次元ベクトルである。

$$\boldsymbol{x_{ij}} = (x_{ij1}, x_{ij2}, \cdots, x_{ijm}), \boldsymbol{c_i} = (c_{i1}, c_{i2}, \cdots, c_{im})$$

このとき、クラスター i の ESS（Error Sum of Squares）を以下のように定義する。

$$ESSi = \sum_{j=1}^{Ni} \| \boldsymbol{x_{ij}} - \boldsymbol{c_i} \|^2$$

ここで、$\| x \|$ はベクトル $\boldsymbol{x}$ の長さ（ユークリッド距離）である。

Ward [2] では、ESS の変化量をそのクラスターが形成されたことに伴う「情報の損失（loss of information）」量であると考え、凝集に伴う「情報の損失」の増加が最も少ないペアを凝集対象として選定することを提案している。

クラスター p と q がクラスター r に凝集する際の「情報の損失」の増加量は以下の通りである。

$$ESSr - (ESSp + ESSq)$$

$$= \sum_{j=1}^{Nr} \| \boldsymbol{x_{rj}} - \boldsymbol{c_r} \|^2 - \left(\sum_{j=1}^{Np} \| \boldsymbol{x_{pj}} - \boldsymbol{c_p} \|^2 + \sum_{j=1}^{Nq} \| \boldsymbol{x_{qj}} - \boldsymbol{c_q} \|^2 \right)$$

$$= \{ Np \cdot Nq \,/\, (Np + Nq) \} \times \| \boldsymbol{c_p} - \boldsymbol{c_q} \|^2 \qquad (2\text{-}1)$$

⑵ 「変数別寄与率算定法」の基本的考え方

m 個の観測データ変数が、「情報の損失」の増加量に対して寄与する度合いを明らかにできれば、クラスター凝集・分離に影響する変数を求めることができる。このために「変数別情報損失寄与率（LIV）」と呼ぶ値を新たに定義する。「情報の損失」の増加量を表す式のなかで、観測データの m 個の変数の値が効いてくるのは「平方距離項」である。式 (2-1) から「平方距離項」だけを抜き出すと、

$$\| \boldsymbol{c_p} - \boldsymbol{c_q} \|^2 = \sum_{i=1}^{m} (c_{pi} - c_{qi})^2$$

ただし、

$$\boldsymbol{c_p} = (c_{p1}, c_{p2}, \cdots, c_{pm}), \boldsymbol{c_q} = (c_{q1}, c_{q2}, \cdots, c_{qm})$$

m 個の変数のうちの特定の変数が「平方距離項」を構成する「重心間距離」に寄与する比率を「変数別情報損失寄与率：LIV（Loss of Information rate by Variable）」と定義し、以下の式 (2-2) で定めよう。n 番目の変数の LIV 値を $LIVn$ と表記する。$\boldsymbol{e_n}$ は変数 n の基底ベクトル、「・」は内積記号である。

$$\begin{aligned} LIVn &= [\{(c_{pn} - c_{qn})\, \boldsymbol{e_n} \cdot (\boldsymbol{c_p} - \boldsymbol{c_q})\} \,/\, \| \boldsymbol{c_p} - \boldsymbol{c_q} \|^2] \times 100 \\ &= \{ (c_{pn} - c_{qn})^2 \,/\, \| \boldsymbol{c_p} - \boldsymbol{c_q} \|^2 \} \times 100 \\ &= \left\{ (c_{pn} - c_{qn})^2 / \sum_{i=1}^{m} (c_{pi} - c_{qi})^2 \right\} \times 100 \end{aligned} \qquad (2\text{-}2)$$

また、LIV 値の全変数総和は、

$$\sum_{n=1}^{m} LIVn = 100$$

である。

この $LIVn$ 値の意味をわかりやすく端的に言えば、「凝集対象クラスターペアの『重心間距離』に対する変数 n の寄与度（%）」である（図 2）。

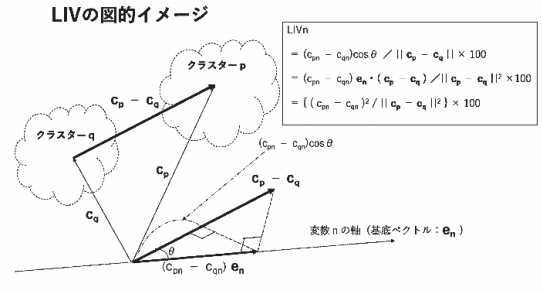

図２　LIV のイメージ

LIVn 値が大きい場合、変数 n が「平方距離項」を大きくする方向に寄与する度合いが高い。つまり、*LIVn* 値が大きな変数 n は、クラスター p とクラスター q を凝集させ難くする方向へ寄与していることになる。別の言い方をすれば、*LIVn* 値が大きい変数 n は、クラスター p とクラスター q を特徴分けする大きな要因である。*LIVn* 値が大きい場合、以下の 5 パターン（以後「*LIV* パターン」と記載）が存在する（$c_{pn} > c_{qn}$ とする）。

- $c_{pn} > c_{qn} > 0$: クラスター p がより強く正の特徴を持つ
- $c_{pn} > c_{qn} \fallingdotseq 0$: クラスター p が正の特徴を持つ
- $c_{pn} > 0 > c_{qn}$: クラスター p が正、クラスター q が負の特徴を持つ
- $c_{pn} \fallingdotseq 0 > c_{qn}$: クラスター q が負の特徴を持つ
- $0 > c_{pn} > c_{qn}$: クラスター q がより強く負の特徴を持つ

これらのパターンの把握は、クラスター間の差異に関する解釈を行う上で役に立つ。なお、ward 法同様、クラスター重心間の平方距離を用いる重心法とメディアン法にも LIV 法は適用可能である。

3.「変数別寄与率算定法」の具体的適用事例

「変数別寄与率算定法」を実際のクラスター分析結果に適用してみよう。実例としては木村・高部 [1] の分析結果を用いる。

図３　木村・高部 [1] の図３

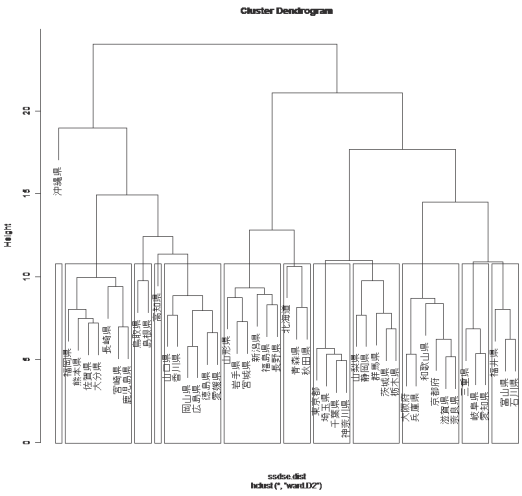

図４　木村・高部 [1] の図２

図 4 は SSDSE-2020C [3] を用いて ward 法によって作成されたデンドログラムである。これは、家計調査データの都道府県庁所在市別データを各都道府県の代表数値と見なし、都道府県を最小単位としている。品目毎の年間世帯平均購入額を都道府県内で比率化し、食料品目毎に標準化（平均 0、標準偏差 1）したデータ（以後「標準化データ A」、$\boldsymbol{x}_{ij}$ ベクトルを行に持つ 47 行×213 列）の分析である。「標準化データ A」と式 (2-2) から簡単に LIV 値を算出することができる。

図４では、千葉（千葉市）と神奈川（横浜市）が最初に凝集する。この時のLIV値を計算し、パレート図（下位側を途中省略。以後同様）にしたものが図５である。

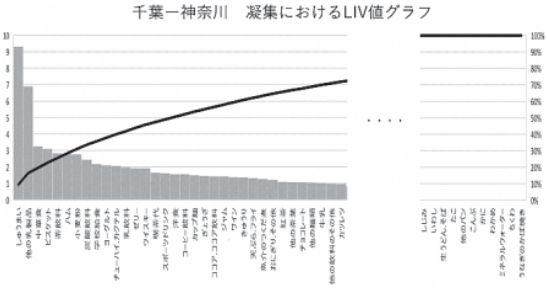

図５　千葉－神奈川　凝集におけるLIV値

横軸は食料213品目名、左縦目盛は食料品目毎のLIV値（棒グラフ）、右縦目盛はLIV累積値（折れ線グラフ）である。213品目中上位17品目のLIV値累計で50%を超えている。

千葉と神奈川の差異を見るためにLIV値の上位５位に着目してみる。「標準化データA」をもとにして「LIVパターン」も見ておく。

	LIV値	「LIVパターン」
しゅうまい：	9.32	神奈川＞千葉＞０
他の乳製品：	6.92	神奈川＞０＞千葉
中華食：	3.26	神奈川＞千葉＞０
ビスケット：	3.09	神奈川＞０＞千葉
茶飲料：	2.85	千葉＞０＞神奈川

１位の「しゅうまい」は千葉も全国平均を上回っているが、神奈川はさらに購入割合が高い。神奈川と千葉の「重心間距離」の9%以上に寄与する食料品目であることを示している。２位の「他の乳製品」とは、「粉ミルク、ヨーグルト、バター、チーズ」以外の乳製品のことである。具体的には「生クリーム、ホイップクリーム、コーヒー・紅茶用のミルク（植物性は除く）」などが含まれる。洋菓子などの材料としての購入であろうか。この品目は千葉が全国平均を下回っているのに対し、神奈川は全国平均を上回っている。３位の外食としての「中華食」は「しゅうまい」同様、千葉も全国平均より多いものの神奈川はさらに上回っている。横浜中華街をかかえる神奈川ならではの結果であろう。逆に千葉を特徴づけるものが５位の「茶飲料」である。小分類「茶類」の「緑茶、紅茶、他の茶葉」といった品目が茶葉に関する品目であるのに対し、「茶飲料」は液体状の茶飲料を示す品目である。ペットボトルの茶類などが含まれる。千葉は全国平均を上回るが神奈川は下回っている。千葉と神奈川が凝集したクラスターでは、これらの差は平均化されることになる。

千葉神奈川クラスターは次に埼玉（さいたま市）と凝集する。この凝集工程におけるLIV値のパレート図が図６である。

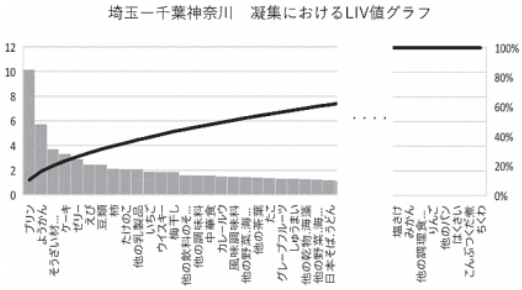

図６　埼玉－千葉神奈川　凝集におけるLIV値

LIV値の上位５位までをリストアップすると、

	LIV値	「LIVパターン」
プリン：	10.15	埼玉＞０＞千葉神奈川
ようかん：	5.74	埼玉＞千葉神奈川≒０
そうざい材料セット：	3.73	埼玉＞０＞千葉神奈川
ケーキ：	3.34	埼玉＞０＞千葉神奈川
ゼリー：	2.90	埼玉＞０＞千葉神奈川

千葉と神奈川を分離していた「しゅうまい」は23位、「他の乳製品」は10位、「中華食」は16位に後退している。これは千葉と神奈川の特徴が平均化された結果である。埼玉では洋和菓子系の品目が好まれることが見て取れる。

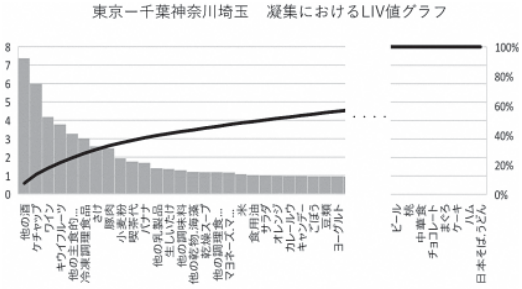

図７　東京－千葉神奈川埼玉　凝集における LIV 値

次に千葉神奈川埼玉クラスターは東京（東京都区部）と凝集する（図 7）。LIV 値計算結果は、

	LIV 値	「LIV パターン」
他の酒：	7.38	東京＞千葉神奈川埼玉≒ 0
ケチャップ：	6.00	千葉神奈川埼玉＞ 0 ＞東京
ワイン：	4.19	東京＞千葉神奈川埼玉＞ 0
キウイフルーツ：	3.79	千葉神奈川埼玉＞ 0 ≒東京
他の主食的外食：	3.28	東京＞千葉神奈川埼玉＞ 0

「ケチャップ」と「キウイフルーツ」は東京の購入率が他の 3 県平均より著しく少ないという結果である。逆に、千葉神奈川埼玉 3 県の食料嗜好の特徴といえる。「他の酒」、「ワイン」は東京での嗜好が高い。南関東 1 都 3 県のクラスターはこのような違いをもとに凝集形成が進んだのである。

次に、山梨（甲府市）と静岡（静岡市）の凝集に関して見てみよう。

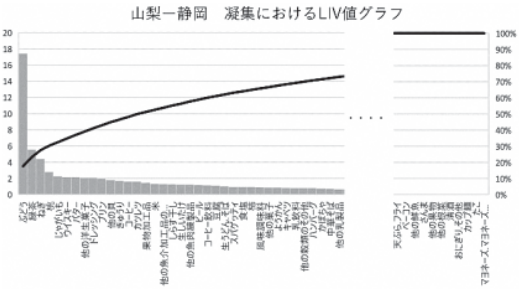

図８　山梨－静岡　凝集における LIV 値

「ぶどう」と「桃」は山梨、「緑茶」は静岡と各県の特産品が主要 LIV である。ここでも、双方の特産品が大きく影響を与えていることがわかる。

LIV 値の上位 5 食料品目は以下のようになる。

	LIV 値	「LIV パターン」
ぶどう：	17.43	山梨＞ 0 ＞静岡
緑茶：	5.57	静岡＞ 0 ＞山梨
ねぎ：	4.40	静岡＞ 0 ＞山梨
桃：	2.78	山梨＞ 0 ＞静岡
じゃがいも：	2.25	静岡＞ 0 ＞山梨

次に、山梨静岡クラスターと茨城栃木群馬クラスターの凝集における LIV 値を見てみる。

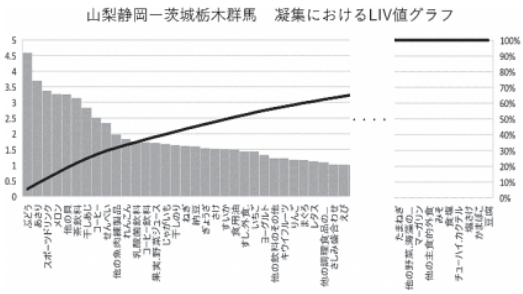

図９　山梨静岡－茨城栃木群馬　凝集における LIV 値

山梨を特徴づけている「ぶどう」など、この段階でもまだ単独県の特産品が影響を残しているものもある。「あさり」は、この工程では山梨静岡クラスターを特徴づける品目として出てきている。山梨と静岡を分ける特徴に「あさり」が登場していないことから山梨と静岡に共通して好まれていることがわかる。「標準化データ A」の値を見てみると、山梨と静岡はともに 2 を超えており解釈が正しいことがわかる。

	LIV 値	「LIV パターン」
ぶどう：	4.59	山梨静岡＞ 0 ＞茨城栃木群馬
あさり：	3.70	山梨静岡＞ 0 ＞茨城栃木群馬
スポーツドリンク：	3.37	茨城栃木群馬＞ 0 ＞山梨静岡
メロン：	3.27	茨城栃木群馬＞ 0 ＞山梨静岡
他の貝：	3.27	茨城栃木群馬＞ 0 ＞山梨静岡
茶飲料：	3.15	茨城栃木群馬＞山梨静岡＞ 0
干しあじ：	2.83	山梨静岡＞茨城栃木群馬＞ 0

最後に、凝集の最終工程を見てみよう。

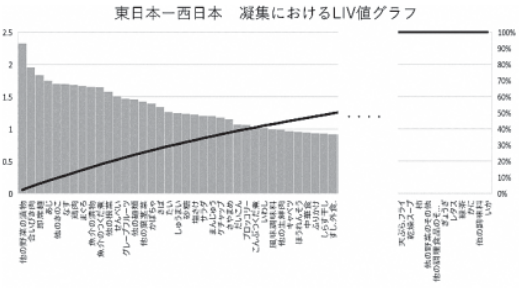

図 10　東日本－西日本　凝集における LIV 値

LIV 値の大きいものから並べてみると LIV 値がかなり僅差であることがわかる。この段階までくると特定の食料品目が距離に大きく寄与することはさすがに無いようだが、「合いびき肉」、「あじ」、「まぐろ」などの、東西を特徴づける品目が登場してくることがわかる。

	LIV 値	「LIV パターン」
他の野菜の漬物：	2.33	東日本＞ 0 ＞西日本
合いびき肉：	1.95	西日本＞ 0 ＞東日本
即席麺：	1.83	西日本＞ 0 ＞東日本
あじ：	1.75	西日本＞ 0 ＞東日本
他のきのこ：	1.70	東日本＞ 0 ＞西日本
なす：	1.70	東日本＞ 0 ＞西日本
鶏肉：	1.69	西日本＞ 0 ＞東日本
まぐろ：	1.67	東日本＞ 0 ＞西日本
魚介の漬物：	1.65	東日本＞ 0 ＞西日本
魚介のつくだ煮：	1.65	東日本＞ 0 ＞西日本
他の根菜：	1.58	東日本＞ 0 ＞西日本
せんべい：	1.51	東日本＞ 0 ＞西日本
グレープフルーツ：	1.47	東日本＞ 0 ＞西日本
他の麺類：	1.46	東日本＞ 0 ＞西日本
他の葉茎菜：	1.43	東日本＞ 0 ＞西日本

4. おわりに

ward 法による階層的クラスター分析結果の解釈を定量的に行う「変数別寄与率算定法（LIV 法）」を提案した。実際に木村・高部 [1] の結果に適用し、クラスター分離に寄与する変数を定量的に把握できることを示した。従来、多変量データの分析結果を読み解くことは困難であったが、着目すべき変数を定量的に絞り込むことが可能となる。

クラスター分析後の解釈手順として、「⑴凝集工程毎の LIV 値」の算出、⑵ LIV 値の大きい変数のリストアップ（着目変数の絞込み）、⑶絞込み変数の LIV パターン分析」という順で行うことが有効だ。LIV 値の算定は、分析前データを Excel で処理して簡単に算出することができる。

分析前データの値にも注意を払う必要がある。山梨の「ぶどう」の例のように、特定の測定データにおける変数値がクラスター全体を引っ張っている場合もある。この場合はクラスターを代表する食料品目として扱うことは不適切である。

以上のように「変数別寄与率算定法（LIV 法）」を用いることにより、結果の解釈においてデータの吟味を適切かつ効率的に実施することができる。

＊参考文献
[1]「家計消費データから見る日本の食料嗜好地域性 ～人文社会系知見との連携も見据えて～」木村敦・高部勲（2021）『ESTRELA』324, pp.36-42, 統計情報研究開発センター .
[2] Joe H. Ward, Jr. (1963) “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, Vol. 58, No. 301(Mar., 1963), pp.236-244.
[3] ㈱統計センター「SSDSE（教育用標準データセット）」https://www.nstac.go.jp/SSDSE/index.html