

クラスタ分析

基本的な考え方と階層的方法

村田 昇

講義の内容

- 第1回：基本的な考え方と階層的方法
- 第2回：非階層的方法と分析の評価

事例

実データによる例

- 総務省統計局より取得した都道府県別の社会生活統計指標の一部
 - 総務省 <https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0>
 - データ https://noboru-murata.github.io/multivariate-analysis/data/japan_social.csv
 - Pref : 都道府県名
 - Forest : 森林面積割合 (%) 2014 年
 - Agri : 就業者 1 人当たり農業産出額 (販売農家) (万円) 2014 年
 - Ratio : 全国総人口に占める人口割合 (%) 2015 年
 - Land : 土地生産性 (耕地面積 1 ヘクタール当たり) (万円) 2014 年
 - Goods : 商業年間商品販売額 [卸売業+小売業] (事業所当たり) (百万円) 2013 年
 - Area : 地方区分

データの概要

分析の目的

クラスタ分析の考え方

クラスタ分析

- クラスタ分析 (**cluster analysis**) の目的
 - 個体の中に隠れている**集まり=クラスタ**を個体間の“距離”にもとづいて発見する方法
- 個体間の類似度・距離 (非類似度) を定義
 - 同じクラスタに属する個体どうしは似通った性質
 - 異なるクラスタに属する個体どうしは異なる性質
- さらなるデータ解析やデータの可視化に利用
- 教師なし学習の代表的な手法の一つ

Table 1: 社会生活統計指標

Pref	Forest	Agri	Ratio	Land	Goods	Area
Hokkaido	67.9	1150.6	4.23	96.8	283.3	1
Aomori	63.8	444.7	1.03	186	183	2
Iwate	74.9	334.3	1.01	155.2	179.4	2
Miyagi	55.9	299.9	1.84	125.3	365.9	2
Akita	70.5	268.7	0.81	98.5	153.3	2
Yamagata	68.7	396.3	0.88	174.1	157.5	2
Fukushima	67.9	236.4	1.51	127.1	184.5	2
Ibaraki	31	479	2.3	249.1	204.9	3
Tochigi	53.2	402.6	1.55	199.6	204.3	3
Gumma	63.8	530.6	1.55	321.6	270	3
Saitama	31.9	324.7	5.72	247	244.7	3
Chiba	30.4	565.5	4.9	326.1	219.7	3
Tokyo	34.8	268.5	10.63	404.7	1062.6	3
Kanagawa	38.8	322.8	7.18	396.4	246.1	3
Niigata	63.5	308.6	1.81	141.9	205.5	4
Toyama	56.6	276.1	0.84	98.5	192.4	4
Ishikawa	66	271.3	0.91	112	222.9	4
Fukui	73.9	216.1	0.62	98.5	167.3	4
Yamanashi	77.8	287.4	0.66	325.3	156.2	4
Nagano	75.5	280	1.65	211.3	194.4	4
Gifu	79	283.7	1.6	192.1	167.9	4
Shizuoka	63.1	375.8	2.91	314.5	211.4	4
Aichi	42.2	472.3	5.89	388.9	446.9	4
Mie	64.3	310.6	1.43	174.3	170.1	5
Shiga	50.5	222.8	1.11	104.9	170.7	5
Kyoto	74.2	267.8	2.05	212.5	196.7	5
Osaka	30.1	216.3	6.96	238.8	451.2	5
Hyogo	66.7	261.2	4.35	197.7	212.5	5
Nara	76.8	207	1.07	182.7	147	5
Wakayama	76.4	251.1	0.76	278.4	136.4	5
Tottori	73.3	249.9	0.45	187.6	162.2	6
Shimane	77.5	214.1	0.55	140.8	141.1	6
Okayama	68	254.8	1.51	184.9	207.8	6
Hiroshima	71.8	286.2	2.24	192.2	304.6	6
Yamaguchi	71.6	216.9	1.11	125.8	158.9	6
Tokushima	75.2	315.4	0.59	313.5	134.5	7
Kagawa	46.4	249.5	0.77	242.9	232.9	7
Ehime	70.3	288.5	1.09	231.6	179.4	7
Kochi	83.3	354.2	0.57	339.9	137.9	7
Fukuoka	44.5	381	4.01	255.6	295.7	8
Saga	45.2	468.7	0.66	230.3	137.9	8
Nagasaki	58.4	428.9	1.08	296	154	8
Kumamoto	60.4	456.6	1.41	285.5	172.5	8
Oita	70.7	360.1	0.92	222.8	148.3	8
Miyazaki	75.8	739.1	0.87	487.7	170.6	8
Kagoshima	63.4	736.5	1.3	351.2	169.4	8
Okinawa	46.1	452.4	1.13	232.8	145.4	8

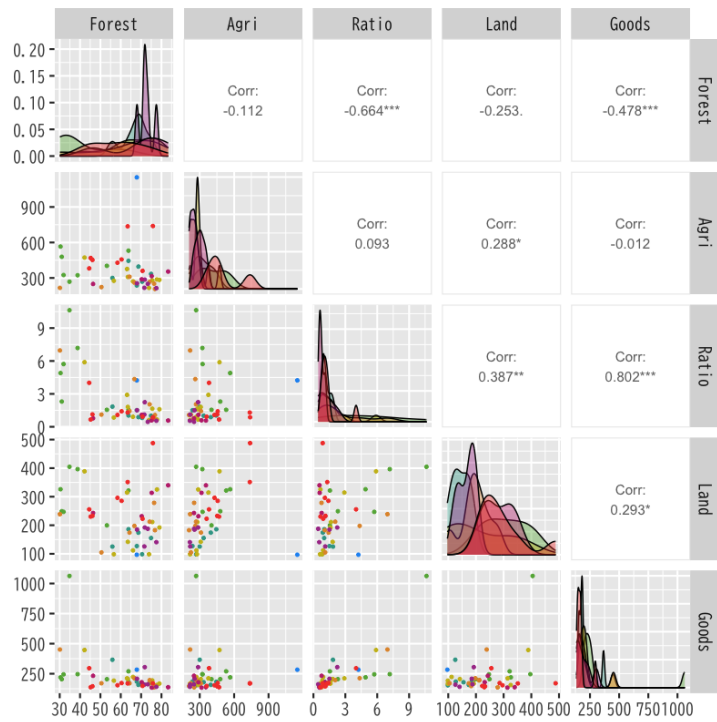


Figure 1: 散布図

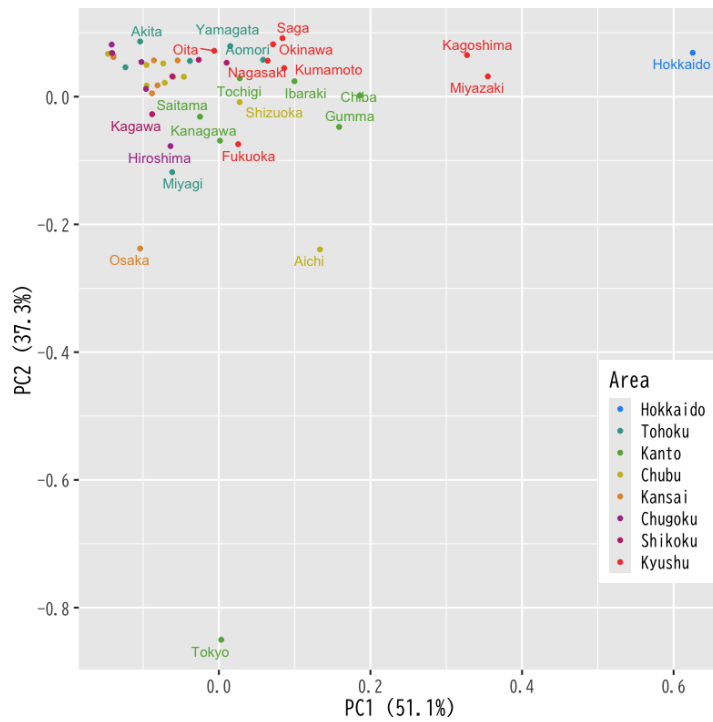


Figure 2: 主成分得点による散布図

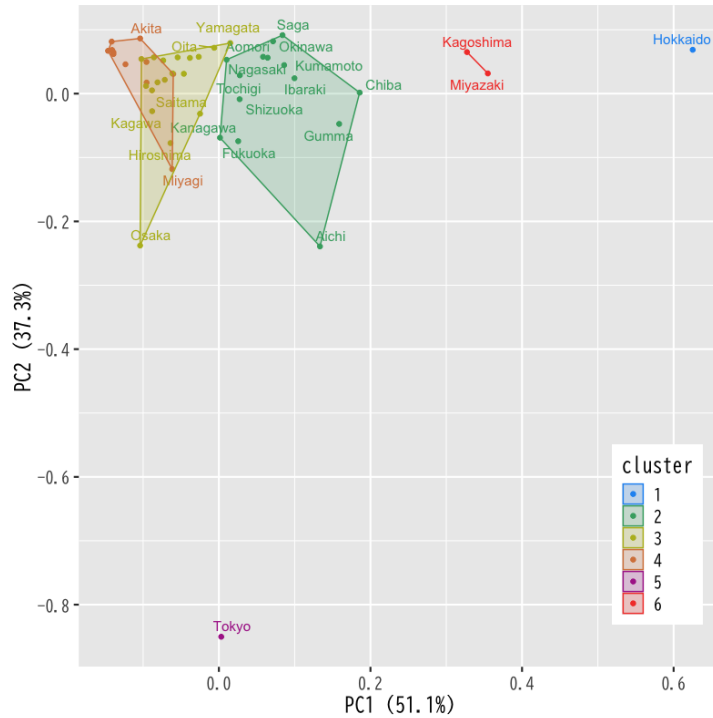


Figure 3: 散布図上のクラスタ構造 (クラスタ分析の概念図)

クラスタ分析の考え方

- 階層的方法
 - データ点およびクラスタの間に **距離** を定義
 - 距離に基づいてグループ化
 - * 近いものから順にクラスタを **凝集**
 - * 近いものが同じクラスタに残るように **分割**
- 非階層的方法
 - クラスタの数を事前に指定
 - クラスタの **集まりの良さ** を評価する損失関数を定義
 - 損失関数を最小化するようにクラスタを形成

階層的方法

凝集的クラスタリング

1. データ・クラスタ間の距離を定義する
 - データ点とデータ点の距離
 - クラスタとクラスタの距離
2. データ点およびクラスタ間の距離を求める
3. 最も近い2つを統合し新たなクラスタを形成する
 - データ点とデータ点
 - データ点とクラスタ
 - クラスタとクラスタ
4. クラスタ数が1つになるまで2-3の手続きを繰り返す

事例

- 社会生活統計指標の一部 (関東)



Figure 4: 凝集的クラスタリング

データ間の距離

データ間の距離

- データ : 変数の値を成分としてもつベクトル

$$\mathbf{x} = (x_1, \dots, x_d)^T, \mathbf{y} = (y_1, \dots, y_d)^T \in \mathbb{R}^d$$

- 距離 : $d(\mathbf{x}, \mathbf{y})$
- 代表的なデータ間の距離
 - Euclid 距離 (ユークリッド ; Euclidean distance)
 - Manhattan 距離 (マンハッタン ; Manhattan distance)
 - Minkowski 距離 (ミンコフスキー ; Minkowski distance)

Euclid 距離

- 最も一般的な距離
- 各成分の差の 2 乗和の平方根 (2 ノルム)

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}$$

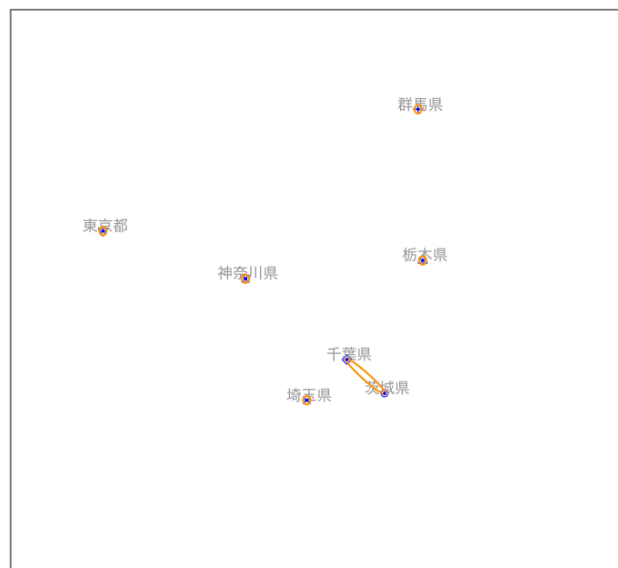


Figure 5: クラスタリングの手続き (その 1)

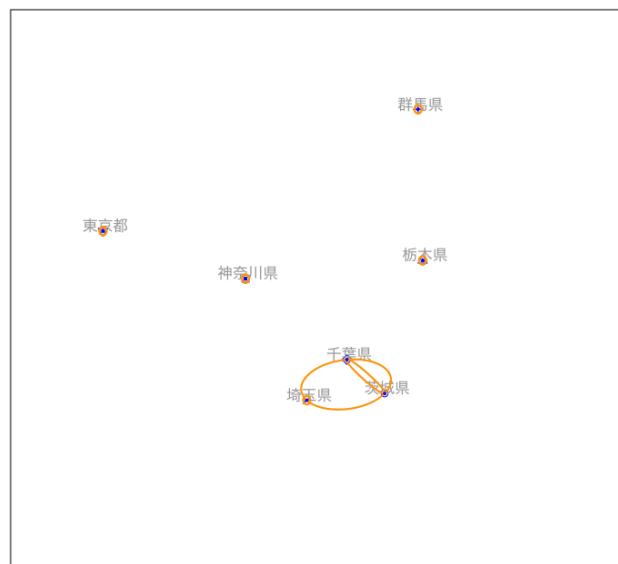


Figure 6: クラスタリングの手続き (その 2)

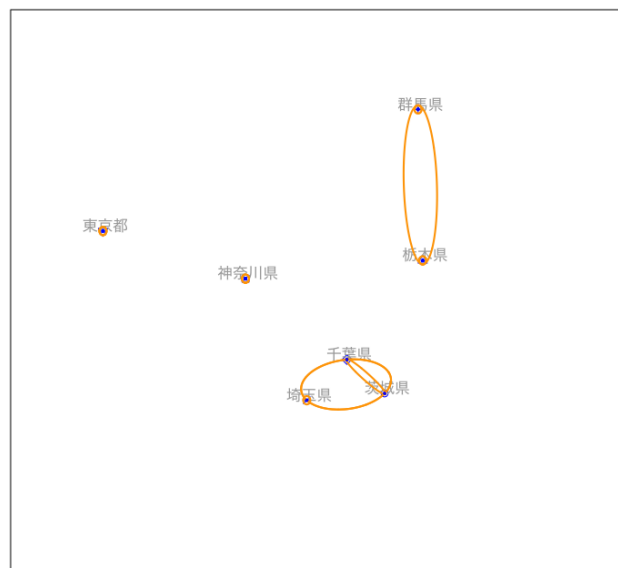


Figure 7: クラスタリングの手続き (その 3)

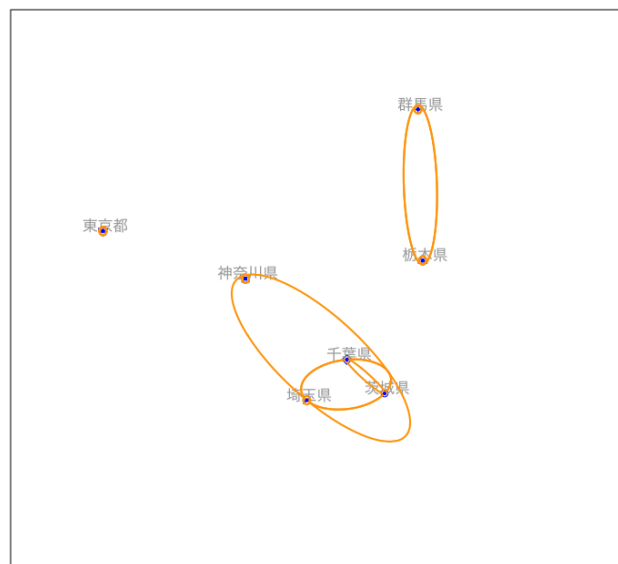


Figure 8: クラスタリングの手続き (その 4)

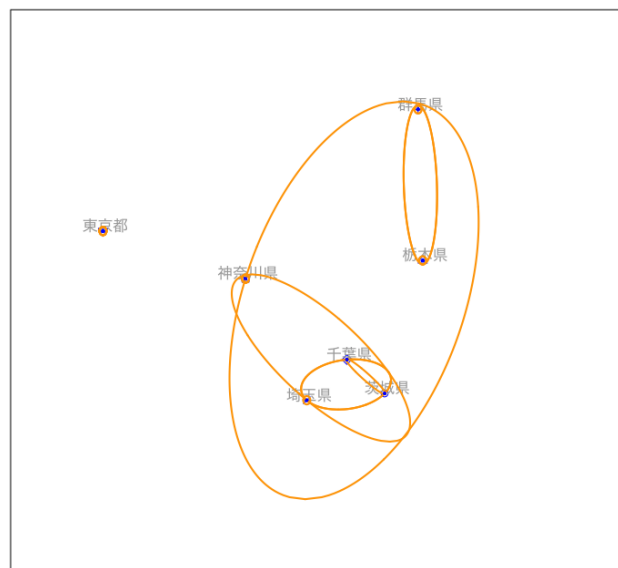


Figure 9: クラスタリングの手続き (その 5)

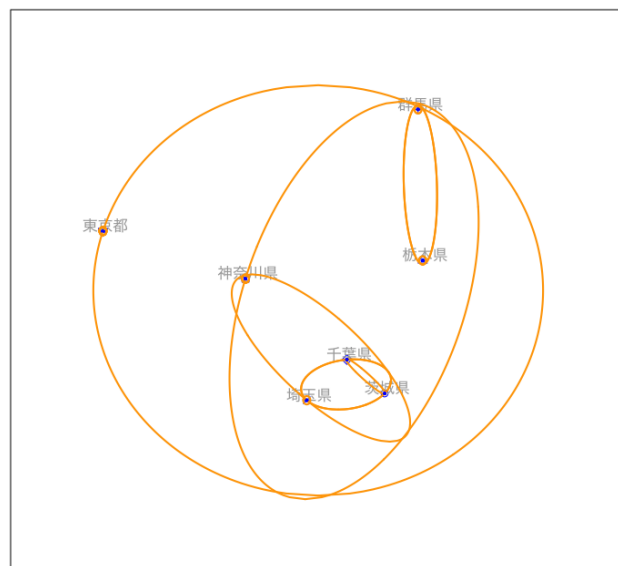


Figure 10: クラスタリングの手続き (その 6)

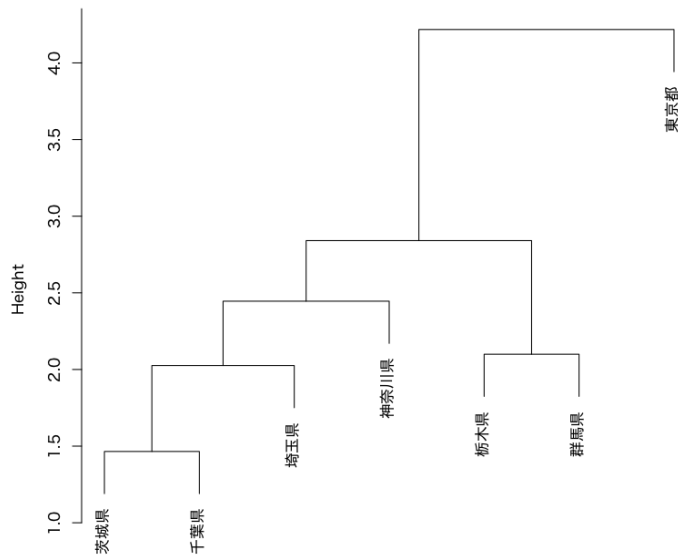


Figure 11: デンドログラムによるクラスタ構造の表示

Manhattan 距離

- 後述する Minkowski 距離の $p = 1$ の場合
- 格子状に引かれた路に沿って移動するときの距離

$$d(x, y) = |x_1 - y_1| + \dots + |x_d - y_d|$$

Minkowski 距離

- Euclid 距離を p 乗に一般化した距離
- 各成分の差の p 乗和の p 乗根 (p -ノルム)

$$d(x, y) = \{|x_1 - y_1|^p + \dots + |x_d - y_d|^p\}^{1/p}$$

その他の距離

- 類似度や乖離度などデータ間に自然に定義されるものを用いることは可能
 - 語句の共起 (同一文書に現れる頻度・確率)
 - 会社間の取引量 (売上高などで正規化が必要)
- 擬似的な距離でもアルゴリズムは動く

演習

問題

- 以下の問に答えなさい

- 距離の定義を述べなさい
- Minkowski 距離において $p \rightarrow \infty$ とするとどのような距離となるか答えなさい

$$d(\mathbf{x}, \mathbf{y}) = \{|x_1 - y_1|^p + \dots + |x_d - y_d|^p\}^{1/p}$$

解答例

- 2変数の実数値関数で以下の3つの条件を満たす
 - 非退化性

$$x = y \Leftrightarrow d(x, y) = 0$$

- 対称性

$$d(x, y) = d(y, x)$$

- 劣加法性 (三角不等式の成立)

$$d(x, y) + d(y, z) \geq d(x, z)$$

- 非負性 $d(x, y) \geq 0$ は3つの条件から自然に導かれる

$$d(x, y) + d(y, x) \geq d(x, x) \quad (\text{劣加法性})$$

$$d(x, y) + d(x, y) \geq d(x, x) \quad (\text{対称性})$$

$$2d(x, y) \geq 0 \quad (\text{非退化性})$$

$$d(x, y) \geq 0$$

- 最大の要素に着目して計算すればよい

$$\begin{aligned} \lim_{p \rightarrow \infty} d(\mathbf{x}, \mathbf{y}) &= \lim_{p \rightarrow \infty} \{|x_1 - y_1|^p + \dots + |x_d - y_d|^p\}^{1/p} \\ &= \lim_{p \rightarrow \infty} \max_k |x_k - y_k| \left\{ \left(\frac{|x_1 - y_1|}{\max_k |x_k - y_k|} \right)^p \right. \\ &\quad \left. + \dots + \left(\frac{|x_d - y_d|}{\max_k |x_k - y_k|} \right)^p \right\}^{1/p} \\ &= \max_k |x_k - y_k| \lim_{p \rightarrow \infty} (1 \text{ 以上の有限値})^{1/p} \\ &= \max_k |x_k - y_k| \end{aligned}$$

- Chebyshev 距離 (最大距離, チェス盤距離) という

- $p \rightarrow -\infty$ の場合は以下となることを確認せよ

$$\lim_{p \rightarrow -\infty} d(\mathbf{x}, \mathbf{y}) = \min_k |x_k - y_k|$$

クラスタ間の距離

クラスタ間の距離

- クラスタ: いくつかのデータ点からなる集合

$$C_a = \{x_i | i \in \Lambda_a\}, C_b = \{x_j | j \in \Lambda_b\}, C_a \cap C_b = \emptyset$$

- 2つのクラスタ間の距離 : $D(C_a, C_b)$
 - データ点の距離から陽に定義する方法
 - クラスタの統合にもとづき再帰的に定義する方法
- 代表的なクラスタ間の距離
 - 最短距離法 (単連結法 ; single linkage method)
 - 最長距離法 (完全連結法 ; complete linkage method)
 - 群平均法 (average linkage method)

最短距離法

- 最も近い対象間の距離を用いる方法

$$D(C_a, C_b) = \min_{x \in C_a, y \in C_b} d(x, y)$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \min\{D(C_a, C_c), D(C_b, C_c)\}$$

最長距離法

- 最も遠い対象間の距離を用いる方法

$$D(C_a, C_b) = \max_{x \in C_a, y \in C_b} d(x, y)$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \max\{D(C_a, C_c), D(C_b, C_c)\}$$

群平均法

- 全ての対象間の平均距離を用いる方法

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{x \in C_a, y \in C_b} d(x, y)$$

- ただし $|C_a|, |C_b|$ はクラスタ内の要素の数を表す

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \frac{|C_a|D(C_a, C_c) + |C_b|D(C_b, C_c)}{|C_a| + |C_b|}$$

距離計算に関する注意

- データの性質に応じて距離は適宜使い分ける
 - データ間の距離の選択
 - クラスタ間の距離の選択
- 変数の正規化は必要に応じて行う
 - 物理的な意味合いを積極的に利用する場合はそのまま
 - 単位の取り方などによる分析の不確定性を避ける場合は平均 0, 分散 1 に正規化
- データの性質を鑑みて適切に前処理

演習

問題

- 以下の問に答えなさい
 - 群平均法におけるクラスタの距離の定義

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{x \in C_a, y \in C_b} d(x, y)$$

から統合前後のクラスタの距離の関係

$$D(C_a + C_b, C_c) = \frac{|C_a|D(C_a, C_c) + |C_b|D(C_b, C_c)}{|C_a| + |C_b|}$$

を導け

解答例

- 定義に従って計算する

$$\begin{aligned} D(C_a + C_b, C_c) &= \frac{1}{|C_a + C_b||C_c|} \sum_{x \in C_a + C_b, y \in C_c} d(x, y) \\ &= \frac{1}{|C_a + C_b||C_c|} \sum_{x \in C_a, y \in C_c} d(x, y) \\ &\quad + \frac{1}{|C_a + C_b||C_c|} \sum_{x \in C_b, y \in C_c} d(x, y) \end{aligned}$$

- (続き)

$$\begin{aligned} &= \frac{|C_a||C_c|}{|C_a + C_b||C_c|} \frac{1}{|C_a||C_c|} \sum_{x \in C_a, y \in C_c} d(x, y) \\ &\quad + \frac{|C_b||C_c|}{|C_a + C_b||C_c|} \frac{1}{|C_b||C_c|} \sum_{x \in C_b, y \in C_c} d(x, y) \\ &= \frac{|C_a||C_c|}{|C_a + C_b||C_c|} D(C_a, C_c) + \frac{|C_b||C_c|}{|C_a + C_b||C_c|} D(C_b, C_c) \\ &= \frac{|C_a|D(C_a, C_c) + |C_b|D(C_b, C_c)}{|C_a| + |C_b|} \end{aligned}$$

解析事例

都道府県別の社会生活統計指標

- 各データを正規化

Forest : 森林面積割合 (%) 2014 年

Agri : 就業者 1 人当たり農業産出額 (販売農家) (万円) 2014 年

Ratio : 全国総人口に占める人口割合 (%) 2015 年

Land : 土地生産性 (耕地面積 1 ヘクタール当たり) (万円) 2014 年

Goods : 商業年間商品販売額 [卸売業+小売業] (事業所当たり) (百万円) 2013 年

- 分析方法: Euclid 距離 + 群平均法

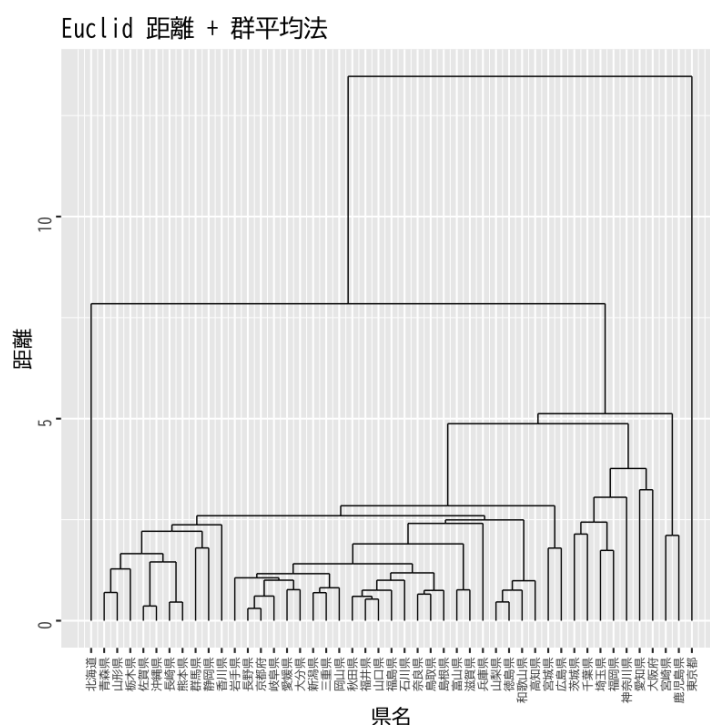


Figure 12: 社会生活統計指標のクラスタ分析 (デンドログラム)

都道府県別好きなおむすびの具

- Web アンケート
 - 「ごはんを食べよう国民運動推進協議会」(平成 30 年解散)
<http://www.gohan.gr.jp/result/09/anketo09.html> (閉鎖)
 - データ <https://noboru-murata.github.io/multivariate-analysis/data/omusubi.csv>
- アンケート概要 (Q2 の結果を利用)

【応募期間】 2009 年 1 月 4 日~2009 年 2 月 28 日

【応募方法】 インターネット、携帯ウェブ

【内 容】

Q1. おむすびを最近 1 週間に、何個食べましたか?

そのうち市販のおむすびは何個でしたか?

Q2. おむすびの具では何が一番好きですか?

A. 梅 B. 鮭 C. 昆布 D. かつお E. 明太子 F. たらこ G. ツナ H. その他

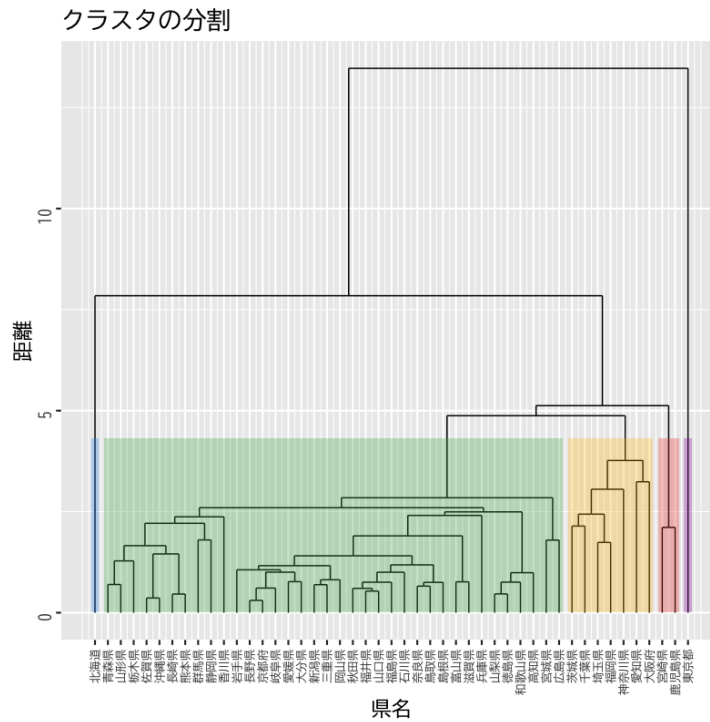


Figure 13: 5分割の例

Q3. おむすびのことをあなたはなんと呼んでいますか？

A. おにぎり B. おむすび C. その他

Q4. おむすびといえば、どういう形ですか？

A. 三角形 B. 丸形 C. 俵形 D. その他

【回答者数】

男性	9,702 人	32.0%
女性	20,616 人	68.0%
総数	30,318 人	100.0%

- 分析方法：Hellinger 距離 (確率分布の距離) + 群平均法

次回の予定

- 第1回：基本的な考え方と階層的方法
- 第2回：非階層的方法と分析の評価

おむすびの具 県別人気アンケート（2009）

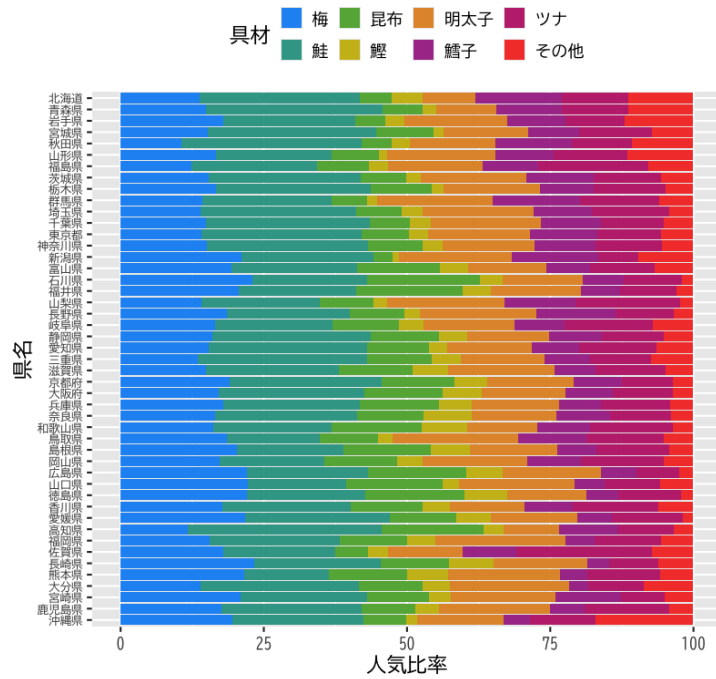


Figure 14: データの概要

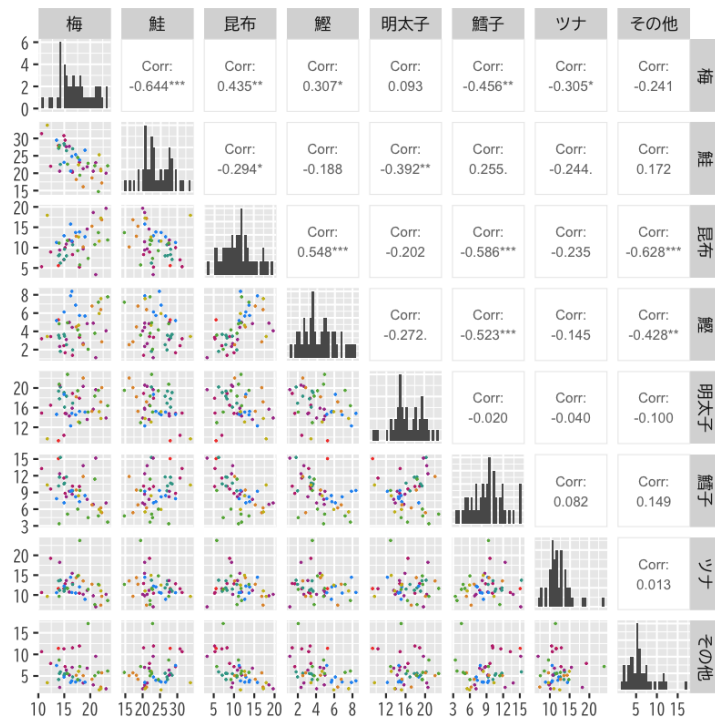


Figure 15: データの散布図

おむすびの具人気アンケートによるクラスタ分析

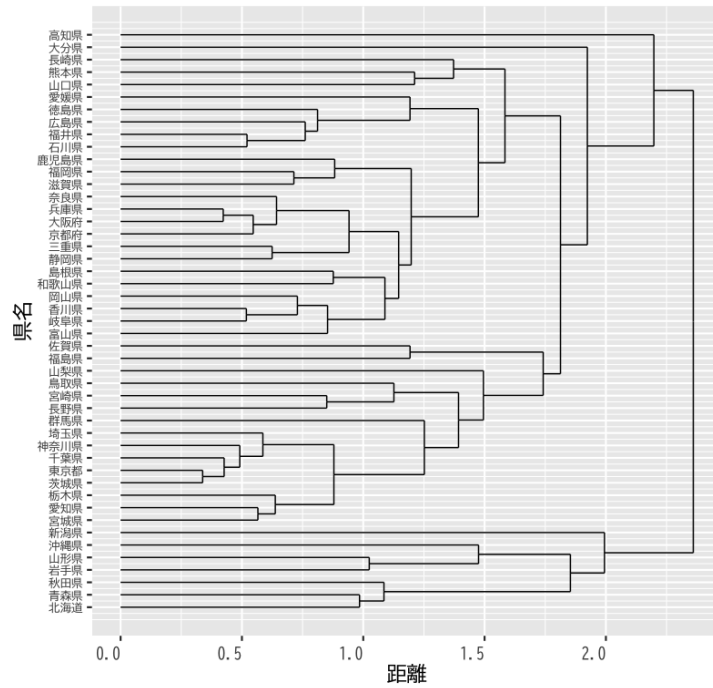


Figure 16: デンドログラム