Clustering of prefectural capitals by food category

Josuke MINAMIGUCHI

1 Background and Motivation

各地域でどの食品の消費量が多いかなどを可視化することで，地域ごとの食文化の多様性を把握しようと考え食品ごとの家計消費量を表すデータを用いた.

具体的には，独立行政法人統計データセンターの SSDSE-家計消費（SSDSE-C) の項目 (`https://www.nstac.go.jp/use/literacy/ssdse/#SSDSE-C`) から都道府県庁所在市別の家計消費データをを取得した. 収録されていたのは,「都道府県庁所在市別、二人以上の世帯の 1 世帯当たり、品目別（食料の全品目）年間支出金額」であった. 集めたデータセットの食品の中分類 (”01 穀類”, ”02 魚介類”, ”03 肉類”, ”04 乳卵類”, ”05 野菜・海藻”, ”06 果物”) によって各県庁所在地をクラスタリングすることでそれぞれの地域の特徴を可視化した.

2 Hypothesis to be verified

食料品の消費データによってそれぞれの地域（県庁所在地）を効率的に分類できるという仮説. また，地理的に近い県庁所在地であれば，同じクラスターに属しやすいと予想した.

3 Method

SSDSE-C-2024.csv から家計消費のデータを読み込んだ. そのうちの，中分類 ”01 穀類”, ”02 魚介類”, ”03 肉類”, ”04 乳卵類”, ”05 野菜・海藻”, ”06 果物” のカラムのみを取得した. 前記の 6 つの食品カテゴリの「二人以上の世帯の 1 世帯当たり、年間支出金額」を特徴量として各県庁所在地をクラスタリングした.

各県庁所在地間の距離はマンハッタン距離，クラスタ間の距離は群平均法により定義した. またデータの標準化は行わなかった. このようにデータを扱ったのは,「ユーグリッド距離 + 標準化なし」,「マンハッタン距離 + 標準化なし」,「ユーグリッド距離 + 標準化あり」,「マンハッタン距離 + 標準化あり」の 4 通りについて凝集係数を比較すると,「マンハッタン距離 + 標準化なし」がもっとも大きい係数だったからである (表 1).

クラスタ数については，クラスタ数 4 ∼ 6 のそれぞれについてシルエット係数を計算し，最も大きいシルエット係数のもの (クラスタ数 $k = 4$) を採用した (Appendix1).

那覇市が外れ値としてクラスタリング結果に影響を与えたかどうかを那覇市を除いたデータフレームについて $k = 3$ のクラスタリングを同様の手法で行うことで結果が変わらないことを確認した (Appdendix2).

表 1: 各クラスタリング手法の凝縮係数

距離	標準化	凝縮係数
ユーグリッド	なし	0.6884225
ユーグリッド	あり	0.6578722
マンハッタン	なし	0.6886454
マンハッタン	あり	0.6811446

4 Results

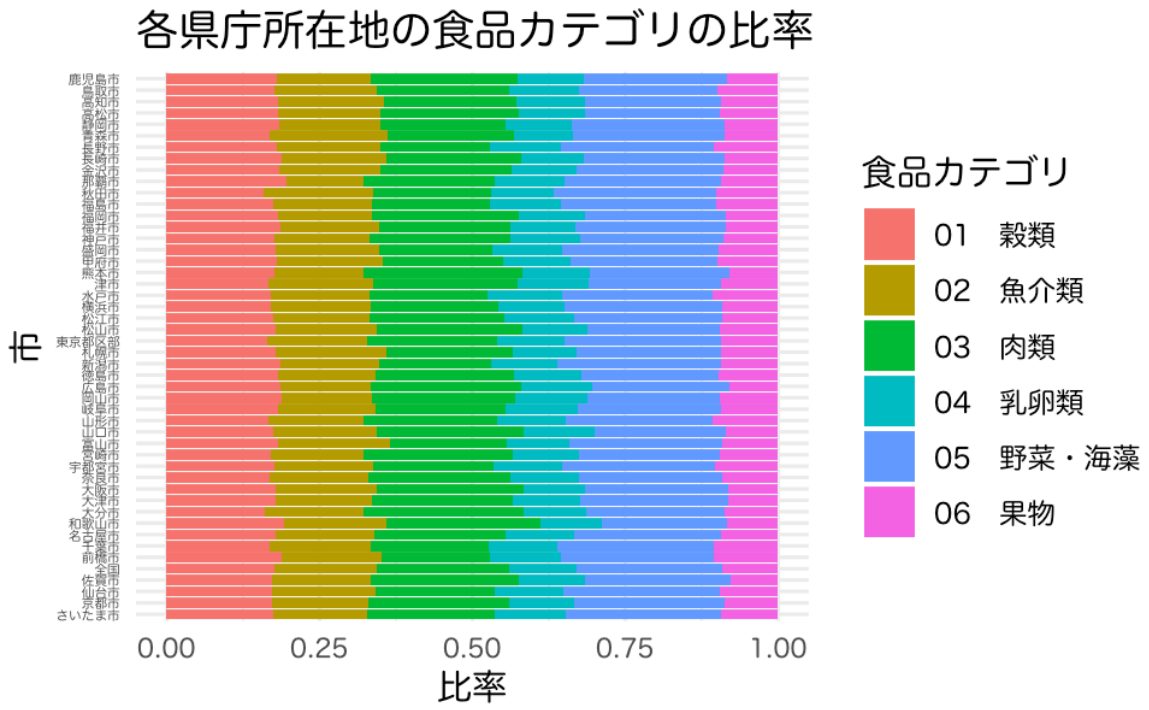

図 1: それぞれの県庁所在地において，各食品カテゴリーに対する支出額の比率を示したグラフ．

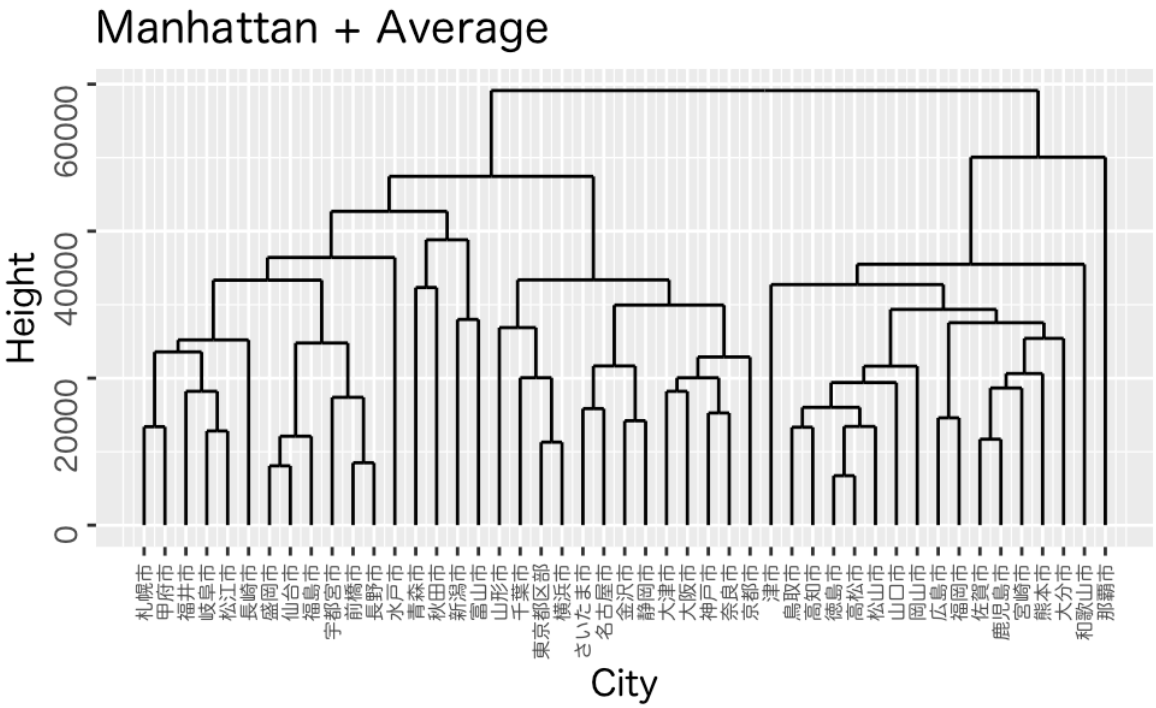

図 2: クラスタリングの結果をデンドログラムで示した.

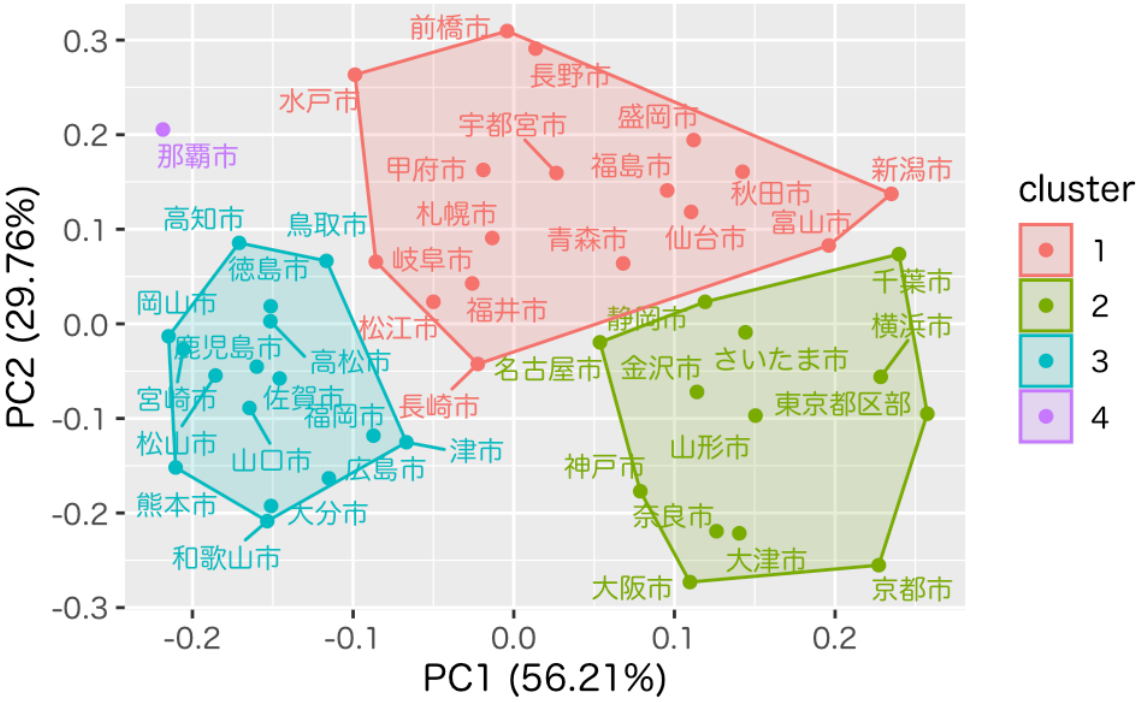

図 3: 第 1 主成分-第 2 主成分のグラフ上にクラスターをプロットした.

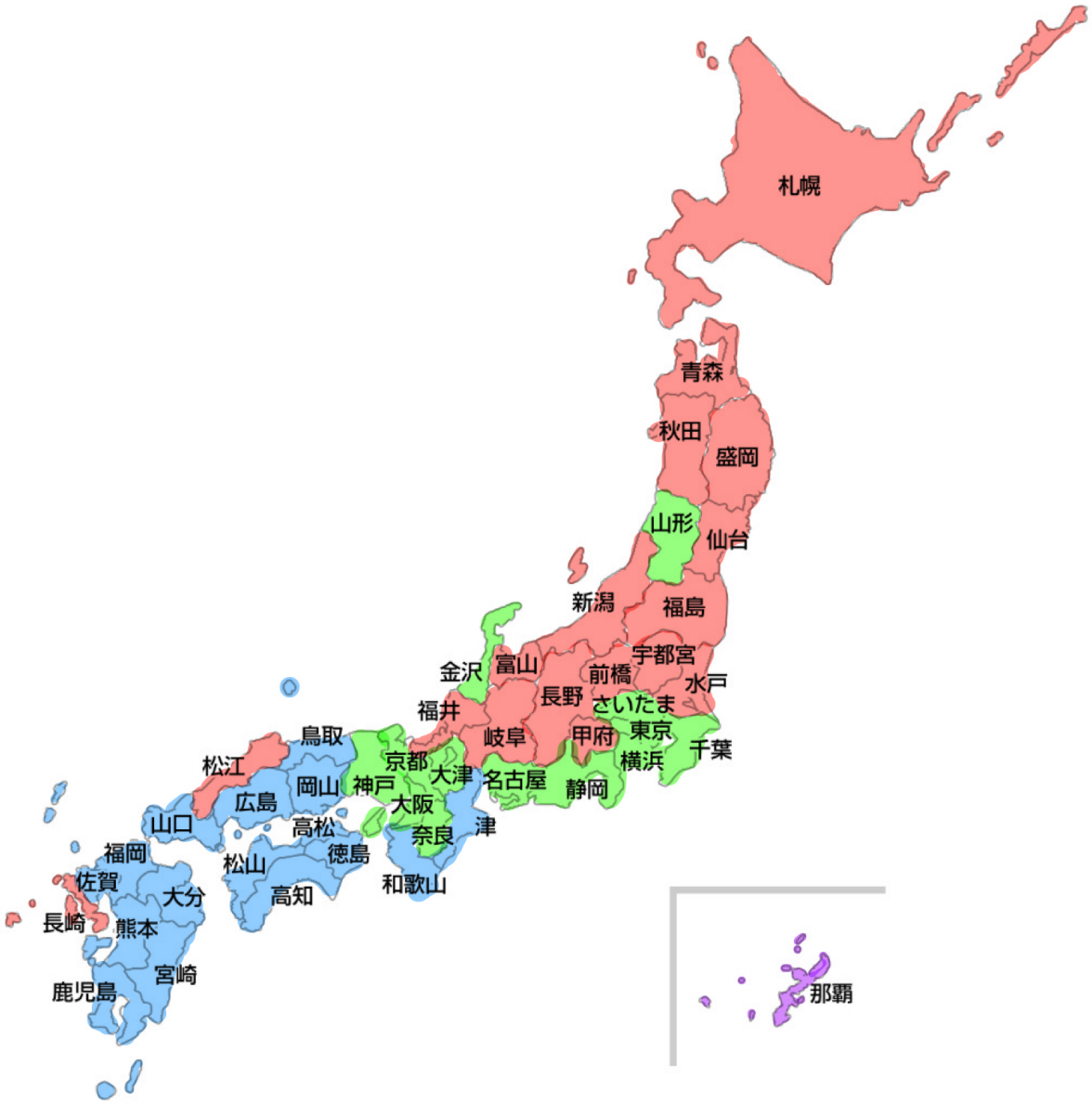

図 4: クラスターの日本地図上での表示

5 Discussion

那覇市は他の県庁所在地とは異なる食生活であることが推測できた.

図 4 から分かるように，クラスタは地理的な近さをよく反映して分布していた．例えば，四国地方，中国地方，九州地方の 3 つではほとんどの県庁所在地がクラスタ 4(青) に属していることが分かった．一方で，一部の地域 (山形市，金沢市，松江市，長崎市) では属するクラスターが集中している地域からは隔絶していた．周囲と異なるクラスタに属するそれらの県庁所在地においては，地理的な要因以外の他の要因が食生活に影響を与えていると予想した.

6 Appendix

6.1 Appendix1

クラスタ数 k を決定するために，$k = 4\,6$ のそれぞれについて，シルエット係数を計算した．その結果 $k = 4$ が適切だと判断した．

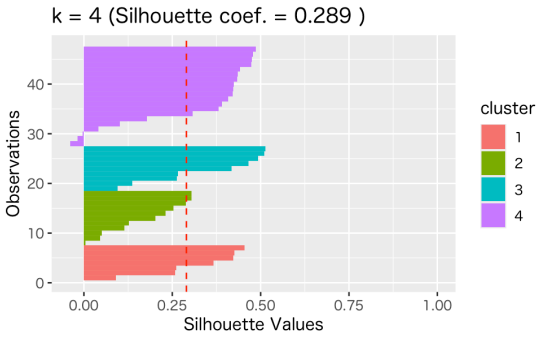

図 5: シルエット係数 (k=4)

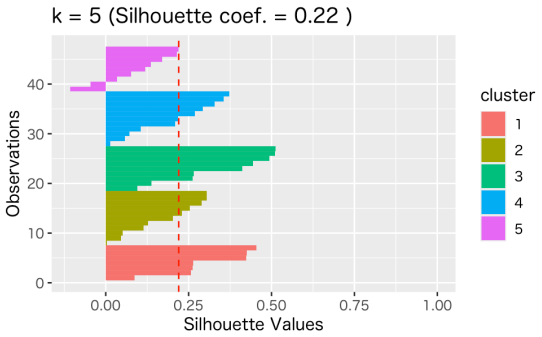

図 6: シルエット係数 (k=5)

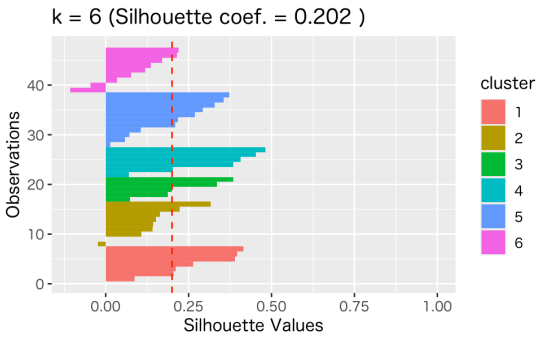

図 7: シルエット係数 (k=6)

6.2 Appendix2

那覇市の単一のクラスタがクラスタリングの結果に影響を与えたかどうかを確認するために，那覇市を除いたデータフレームについて $k = 3$ として同様のクラスタリング (マンハッタン距離＋標準化なし＋群平均法) を行った．那覇市を除いても結果は変わらなかった．

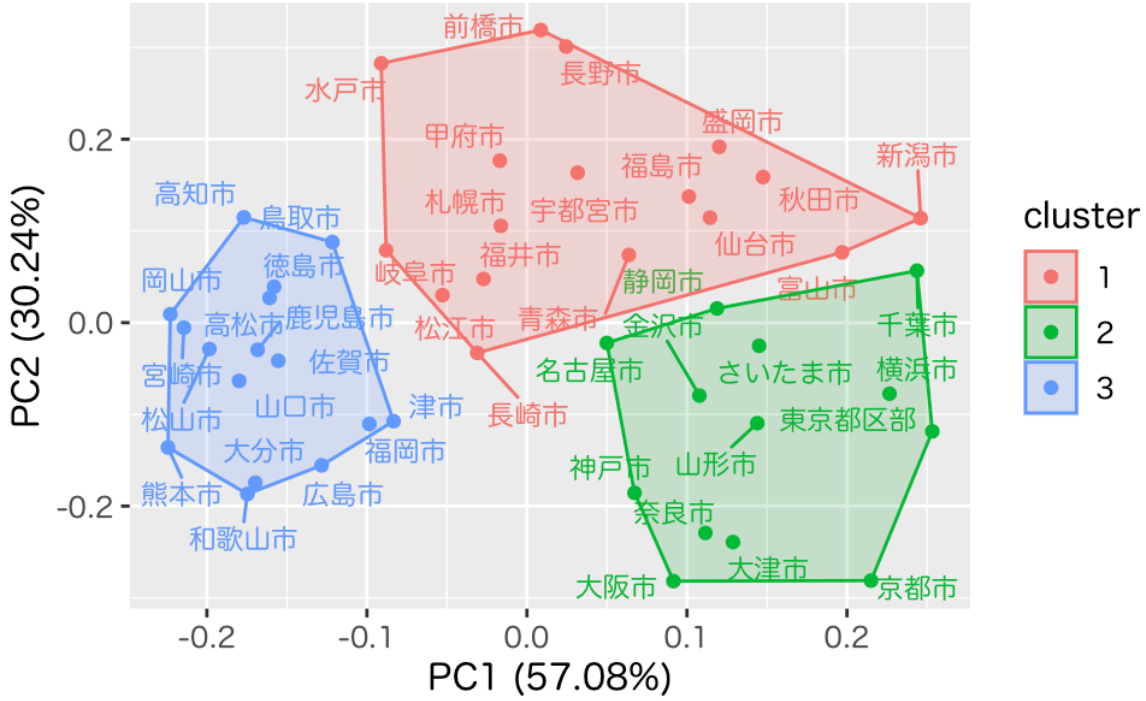

図 8