家計調査データを分析すると
日本の食料嗜好地域性が見えてくる

独立行政法人　統計センター　　理事　木村　敦

＜階層的クラスター分析＞

・高等学校学習指導要領（平成30年告示）解説（情報編）

(3)「情報とデータサイエンス」イ（イ）p51

・高等学校情報科「情報Ⅱ」教員研修用教材（文部科学省）

16「クラスタリングによる分類」

1「階層的クラスタリング」p152

用いる多変量データと分析の事例

- スマートホンが広く普及し、いつでもどこでも情報にアクセスできる時代。Amazonや楽天などを使えば、欲しいものがいつでもどこでも手に入る時代。仕事もテレワーク。今まさに、エリア・ロケーションフリーな時代です。

- しかし、日常的な食の嗜好についてはこの時代においても地域性が残っていることが良く知られています。（このおかげで「秘密のケンミンSHOW」も成立する。）

- 総務省統計局の家計調査*データを多変量分析することによって、都道府県を最小単位とした食の地域性を明らかにしてみましょう。

＊ 家計調査は、国民生活における家計収支の実態を把握し、国の経済政策・社会政策の立案のための基礎資料を提供することを目的に総務省統計局が毎月実施している基幹統計の一つ。調査対象世帯に、日々の家計上の収入および支出を家計簿に記入し、提出してもらう方法で調査を実施。

多変量分析の一つ「クラスター分析」が役立ちます

クラスター分析：多変量データをデータ間の類似性によりグループ分けする手法。大きく分けると以下の２種類に分けられます。

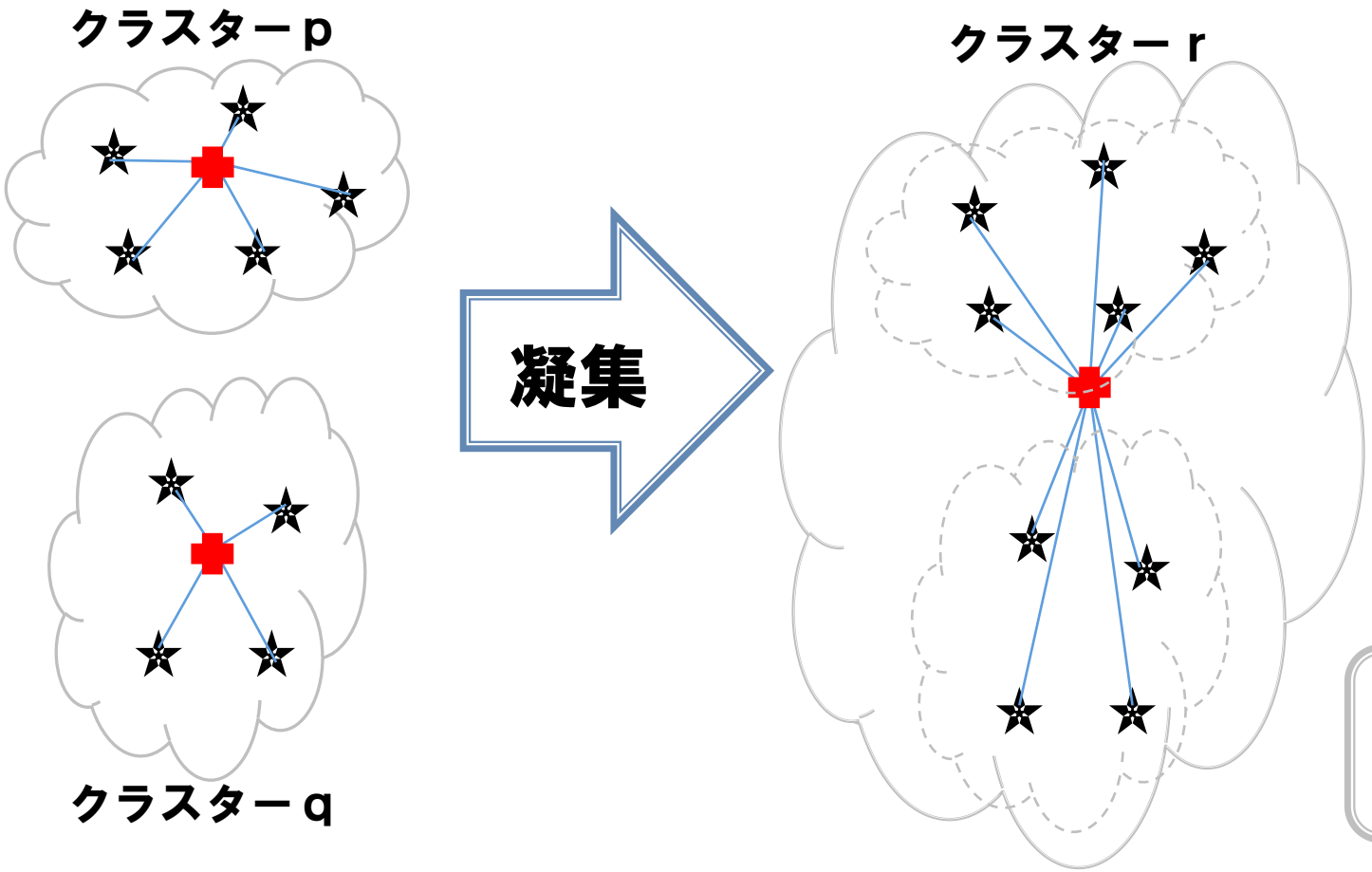
- 階層的クラスター分析
  - 最終的に分類するクラスターの数が未定な場合。
  - 凝集のプロセスに興味がある場合。
  - データ量がさほど大きくない場合。（計算量が発散しない程度。）
  - 古典的な手法とされる。手法例：群平均法、重心法、 ウォード法(ward法) など

- 非階層的クラスター分析
  - クラスター数があらかじめ決まっている場合。
  - 凝集のプロセスには興味がない場合。
  - データ量が極めて多く、階層的クラスター分析が計算量的に不可能な場合。
  - さらにハードクラスタリングとソフトクラスタリングに分けられる。
    手法例：ハード：k平均法(k-means法)など、ソフト：Fuzzy c-meansなど

今回は、階層的クラスター分析のなかのウォード法を使った事例を紹介します。

ウォード法による凝集アルゴリズムの概要

①クラスター生成に伴う「情報の損失(ESS)」:
　クラスター内の各要素とクラスター重心間の平方ユークリッド距離の総和で定義。
②クラスター間の「非類似度（情報の損失の増加量）」：　ESSｒ－（ESSｐ＋ESSｑ）
③「非類似度」が最小となるクラスターpとqを凝集ペアに選定し、凝集を繰り返す。



「教育用標準データセットSSDSE-C（家計調査）」を使う

家計調査データを教育用に使いやすく加工したもので、統計センターが作成・公表。
（47都道府県庁所在市　×　213食料品目、世帯当たり年間購入額の３年間平均値）

・縦（行の数）：項目情報（2）＋全国・都道府県庁所在市（48）＝ 50 行

・横（列の数）：地域情報（3）＋データ項目（227）＝ 230 列

年次情報

項目情報：データ項目の項目名／データ項目のコード

全国＋47県庁所在市

2017_2019	Prefecture	City	LA03	LB00	LB01	LB011001	LB012001
地域コード	都道府県	市	世帯人員	食料（合計）	01　穀類	米	食パン
R00000	全国	全国	2.98	954715	77480	23736	9461
R01100	北海道	札幌市	2.96	910399	81474	30994	8496
R02201	青森県	青森市	2.98	878930	71992	23773	7777
∫	∫	∫	∫	∫	∫	∫	∫
R45201	宮崎県	宮崎市	2.92	846584	65809	21411	7153
R46201	鹿児島県	鹿児島市	2.80	886950	70929	23991	7550
R47201	沖縄県	那覇市	3.02	789143	68947	27643	7419

地域コード　都道府県　県庁所在市

地域情報

～	LB121110	LB121201	LB121202	LB122001
～	他の主食的外食	喫茶代	飲酒代	学校給食
～	54404	7005	18785	10081
～	22506	5665	26189	15619
～	13222	2571	14178	10745
	∫	∫	∫	∫
～	48246	3318	20452	11091
～	40733	3773	16663	12428
～	46441	3291	19436	10483

データ項目（227）

都道府県庁所在市のデータですが、今回は都道府県を代表する値としてみなすこととします。

SSDSE-C-2020の簡単な前処理

・縦（行の数）： 項目情報（2） ＋ 全国・都道府県庁所在市（48） ＝ 50 行

・横（列の数）： 地域情報（3） ＋ データ項目（227） ＝ 230 列

①不要行列削除

年次情報

項目情報 { データ項目の項目名 / データ項目のコード

2017-2019	Prefecture	City	LA03	LB00	LB01	LB011001	LB012001
地域コード	都道府県	市	世帯人員	食料（合計）	01 穀類	米	食パン
R00000	全国	全国	2.98	954715	77480	23736	9461
R01100	北海道	札幌市	2.96	910399	81474	30994	8496
R02201	青森県	青森市	2.98	878930	71992	22772	7777
∫	∫	∫	∫	∫	∫	∫	∫
R45201	宮崎県	宮崎市	2.92	846584	65809	21411	7153
R46201	鹿児島県	鹿児島市	2.80	886950	70929	23991	7550
R47201	沖縄県	那覇市	3.02	789143	68947	27643	7419

全国＋47県庁所在市

地域コード　都道府県　県庁所在市

地域情報

～	LB121110	LB121201	LB121202	LB122001
～	他の主食的外食	喫茶代	飲酒代	学校給食
～	54404	7005	18785	10081
～	22506	5665	26[illegible]89	15619
～	18222	2571	14[illegible]78	10745
	∫	∫	∫	∫
～	48246	3318	20[illegible]52	11091
～	40733	3773	16[illegible]63	12428
～	46441	3291	19[illegible]36	10483

データ項目（227）

②行で比率化

③列で標準化（平均０、標準偏差１）

処理後のデータ（以後「標準化データA」） 47行×213列

ssdse_std.csv - Excel

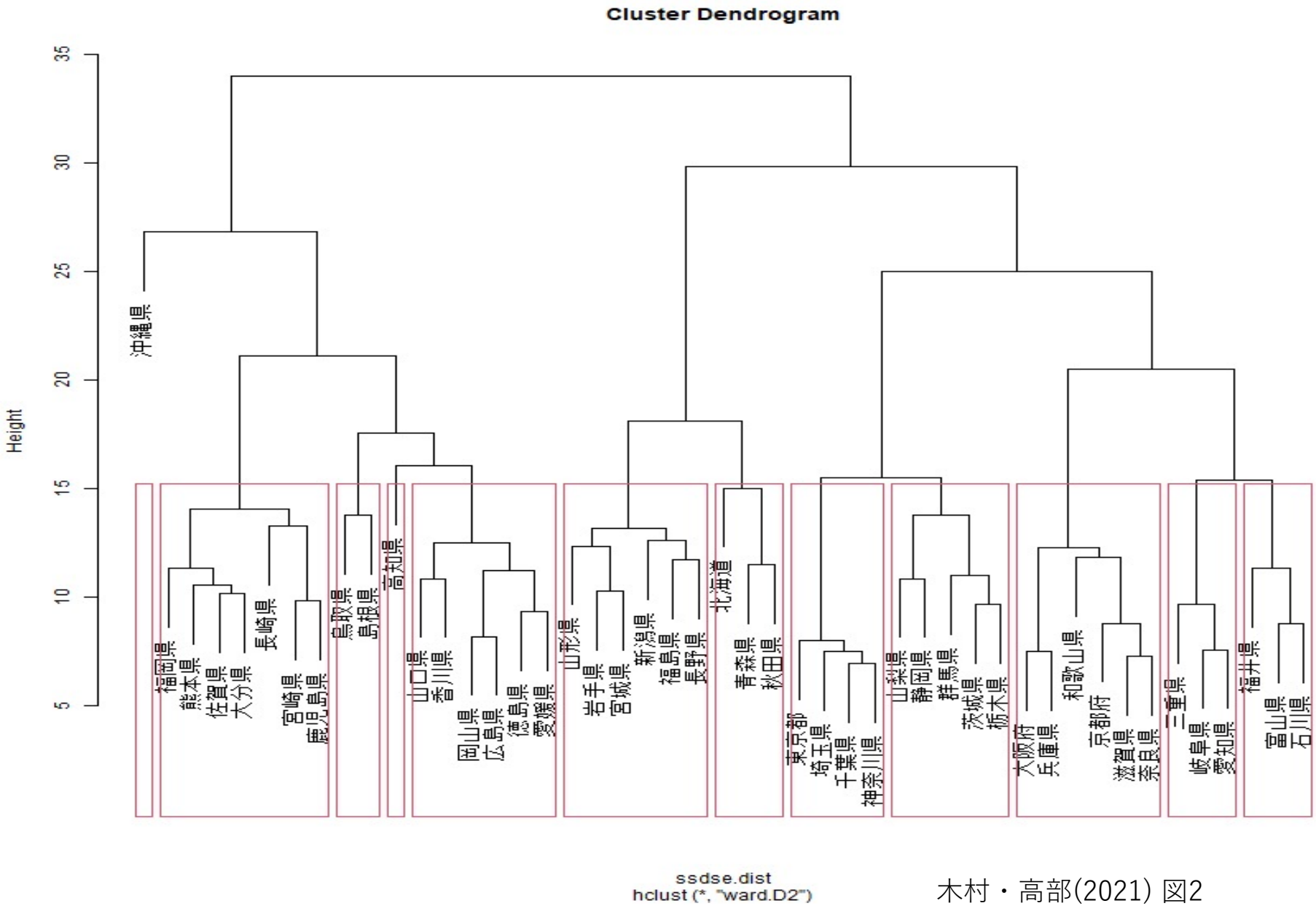
ファイル　ホーム　挿入　ページ レイアウト　数式　データ　校閲　表示　Power Pivot　実行したい作業を入力してください...　共有

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1		米	食パン	他のパン	生うどん.そ	乾うどん.そ	スパゲッテ	中華麺	カップ麺	即席麺	他の麺類	小麦粉	もち	他の穀類の	まぐろ	あじ
2	北海道	2.93945	-0.44534	-0.66078	-0.26615	-0.34889	2.26504	0.50922	1.13311	-0.78924	0.82054	0.48033	0.52788	1.00451	0.00651	-1.5
3	青森県	0.79537	-0.79279	-1.11417	-0.394	-0.13969	-0.40134	2.43213	2.98065	0.50975	-0.55196	-0.72201	-0.69488	-1.86799	0.56841	-0.7
4	秋田県	-0.85599	-1.99537	-1.01526	0.02381	3.22276	-1.00638	1.97977	1.60412	-0.18499	0.28346	-0.2795	-1.16353	-2.281	0.49063	-0.6
5	岩手県	0.84045	-0.90318	-0.29453	-0.14977	0.25099	0.38533	3.12153	1.63011	-0.12019	1.72733	0.13872	-1.098	-0.40732	0.78862	-1.0
6	宮城県	-1.03196	-1.17336	-1.07661	-0.47075	0.3953	0.68454	1.56024	0.88385	-0.45014	1.33874	-0.18148	-0.76654	-0.83268	0.66387	-0.8
7	山形県	0.74056	-1.53069	-1.48861	1.23347	1.766	0.35459	1.37848	0.82243	-0.9098	0.54647	-0.77818	0.03128	-1.72227	0.34088	-1.1
8	福島県	0.34863	-1.86266	-1.353	-0.72333	0.69728	-1.34084	0.23204	1.52305	-0.81966	1.77094	-0.16192	-0.68571	-1.48408	0.95238	-1
9	新潟県	0.60352	0.31931	0.00233	-0.96365	1.88509	0.91885	0.1452	1.63363	0.37999	-0.21113	0.718	1.14504	-0.84333	-0.37764	-0.5
10	長野県	-0.11257	-0.86263	-1.34182	1.31156	0.89906	0.9826	0.67096	-0.04979	-0.74053	1.46436	3.43941	-0.59738	-0.65499	0.4344	-0.8
11	茨城県	-1.12664	-0.60281	-1.4349	-0.16167	0.15219	-0.47884	0.00096	0.42712	-0.9889	0.30247	-0.66235	0.01563	-0.81926	0.45286	-0.9
12	栃木県	-0.58599	-0.5269	-1.24842	0.08249	0.09416	0.18902	-0.00321	0.60616	-0.69222	0.0663	-0.33296	0.35533	-0.16686	1.63464	-0.9
13	群馬県	0.9074	-1.19795	0.65896	1.51041	-0.56115	-0.00675	-0.09529	0.51485	-1.07667	-0.20439	0.3335	-0.71581	-0.49912	1.27747	-0.5
14	山梨県	-0.08297	-0.56328	-0.14359	0.65079	-0.63434	-1.19293	0.5907	-0.13715	-1.31209	0.57397	-1.14155	-0.59729	-1.05863	1.94965	-0.4
15	静岡県	1.67602	-0.72695	-0.41663	-0.89763	-0.17125	0.30969	0.99746	-0.12885	-0.82526	1.57007	-0.56028	-0.09384	0.32139	3.01768	0.0
16	埼玉県	-0.17251	-0.68327	-0.15889	0.14193	-0.5999	0.95878	-0.20858	-0.67129	-1.18287	1.74187	-0.22511	-0.32995	0.66289	0.85806	-0.2
17	千葉県	-0.76113	-0.11134	-0.15362	-0.60876	-0.39612	-0.13955	0.04353	-0.50927	-0.99432	1.06149	-1.38255	0.29777	0.01299	1.03867	-0.1
18	東京都	-1.38157	-0.07679	-0.10024	-1.02098	-0.34412	0.13298	-0.6527	-1.56417	-1.52524	1.31238	-1.73788	-0.02843	-0.60068	1.07068	-0.3
19	神奈川県	-0.31796	0.06923	-0.16725	-0.63481	-0.13881	0.24748	-0.52456	-1.71441	-1.31697	2.07366	0.25791	0.50382	0.21234	1.25758	0.0
20	富山県	-0.45292	0.07032	-0.43391	-0.25925	-0.02071	-1.48605	-0.13996	0.88396	-0.19305	-1.26316	-1.65837	2.91521	-0.91697	-0.11996	0.5
21	石川県	-0.79884	0.53391	-0.1715	-0.11131	-0.16594	-0.69307	0.9569	-0.80732	-0.84559	-0.3863	-0.53451	1.96949	-1.35198	-0.70804	0.4
22	福井県	1.19658	-0.45167	0.35561	0.75486	-0.06085	-0.57276	-0.09792	-0.91836	-0.63032	-1.25827	-1.63429	3.30897	-0.97676	-0.81428	-0.4
23	岐阜県	0.31239	0.01114	-0.44597	-0.3446	-0.24087	-0.22587	-0.40079	-0.43414	-0.3895	0.37369	-0.60477	1.375	0.02142	0.26375	-0.6
24	愛知県	0.26132	0.67612	-0.46669	0.67986	-0.59969	0.55548	0.01463	-0.32254	-0.15334	0.24453	-0.68852	0.82443	0.35841	0.53929	-0.7

ssdse_std

準備完了　130%

「標準化データA」のクラスター分析（ward法）結果



木村・高部(2021) 図2

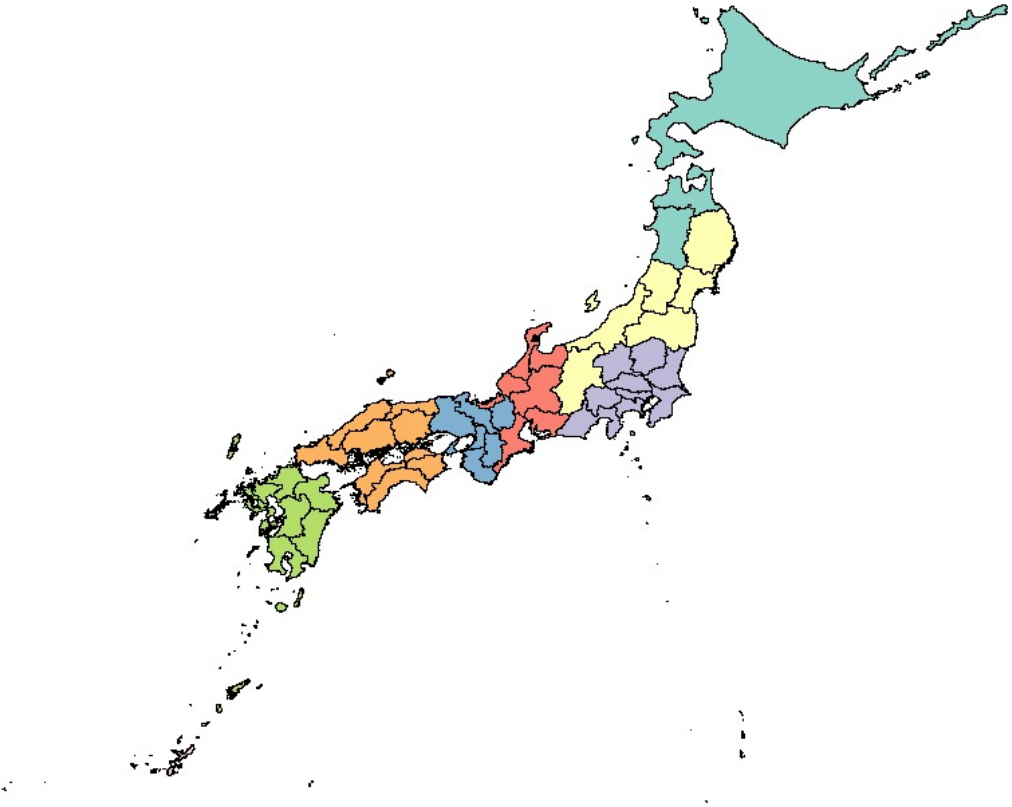
（参考）分析プログラム(R言語で16行)

```
###############    階層型クラスター分析スクリプト（エストレーラ2021年3月号の「研究ノート」用）   ################

ssdse_2020c <- read.csv(“SSDSE-2020C.csv”)                                                    # csvファイル読込

##################################    前処理部    ################################
library(dplyr)
ssdse_2020c_small <- select(ssdse_2020c, -X2017_2019, -City, -LA03, -LB01, -LB02,
                                    -LB03, -LB04, -LB05, -LB06, -LB07, -LB08, -LB09, -LB10, -LB11, -LB12)     #
ssdse_2020c_small_pref <- filter(ssdse_2020c_small, Prefecture != "全国")                      #
colnames(ssdse_2020c_small_pref) <- ssdse_2020c_small_pref[1,]                                 #
ssdse_2020c_edited_pref <- filter(ssdse_2020c_small_pref, 都道府県!=“都道府県”)                 # ① 不要行列削除
rownames(ssdse_2020c_edited_pref) <- c(ssdse_2020c_edited_pref$都道府県)                        #
ssdse <- select(ssdse_2020c_edited_pref, -都道府県)                                             #
write.csv(ssdse, file=“ssdse_2020c_edited.csv”, quote = FALSE)                                 #
ssdse <- read.csv("ssdse_2020c_edited.csv", row.name = 1)                                      #

ssdse.p <- (ssdse %>% select(米:学校給食) %>% sweep(.,1,c(as.numeric(ssdse[,1])),“/”)*100)        # ② 行での比率化
ssdse.std <- scale(ssdse.p)          #「標準化データA」完成！(ssdse.std)                        # ③ 列での標準化

################################    クラスター分析部    ############################
ssdse.dist <- sqrt(1/2)* dist(ssdse.std)                                                       # 非類似度行列の作成
ssdse.hclust <- hclust(ssdse.dist, method=“ward.D2”)                                           # ウォード法クラスタ分析

############################    デンドログラム描画部    #########################
plot(ssdse.hclust)                                                                             # 樹形図の表示
rect.hclust(ssdse.hclust, k=12)                                                                # クラスター赤枠描画
```

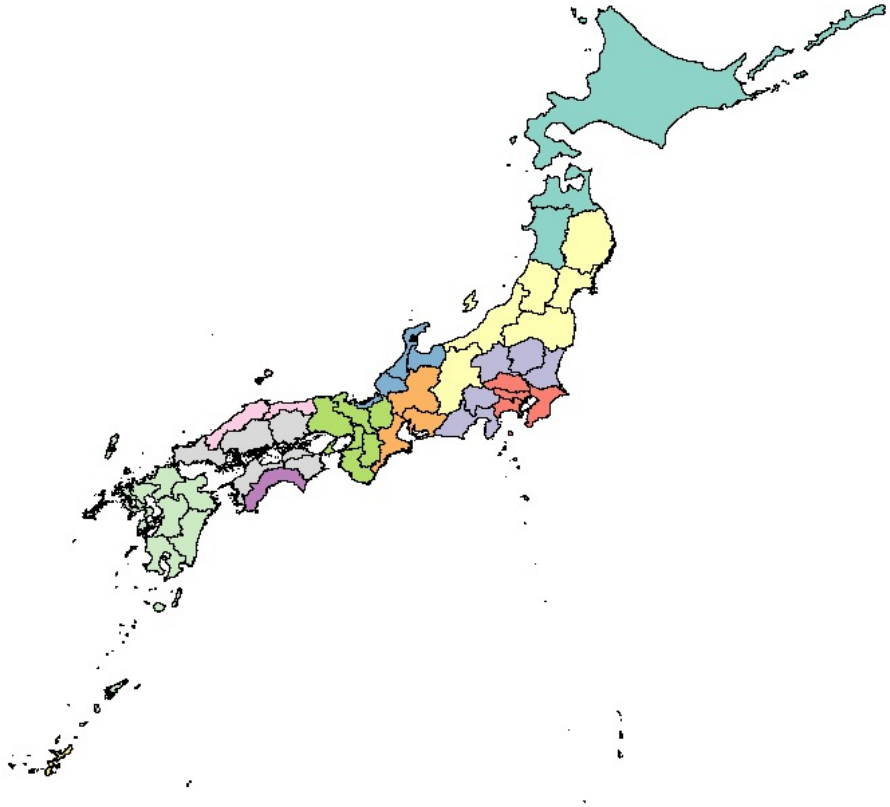
デンドログラムを日本地図にしてみると

家計調査データには地理的情報は全く含まれていないにも関わらず、地理的近接クラスターが現れました！？

クラスター数８の場合
木村・高部(2021) 図4



クラスター数１２の場合
木村・高部(2021) 図3



そこで、分析結果の評価をしましょう

妥当性：・分析に使っていない情報や既存の知見から評価してみる。

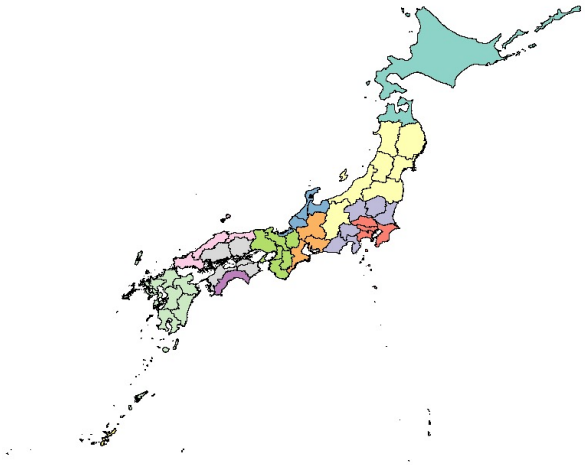
→ 一般常識的には、食の嗜好は地理的に近接したエリアで類似性が高い。今回の分析結果を支持する。

安定性：・異なる調査データ（年度の異なるデータなど）を用いた同様な分析の結果を比較してみる。

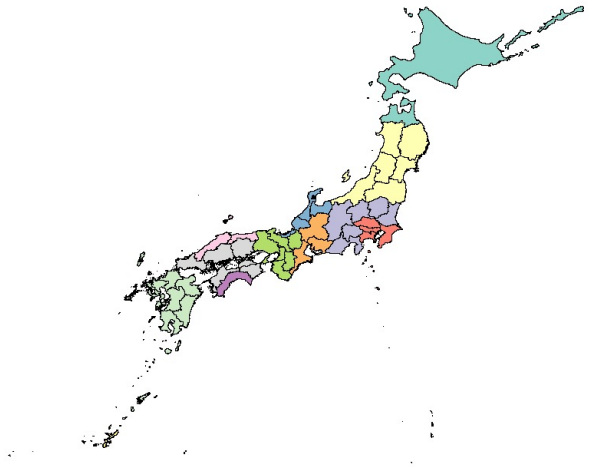
→ 幸いなことに、総務省統計局の家計調査は、2007年から継続的に県庁所在市データが利用可能。e-Stat（政府統計の総合窓口）から入手できます。

安定性：異なる調査年次の「家計調査データ」の分析結果

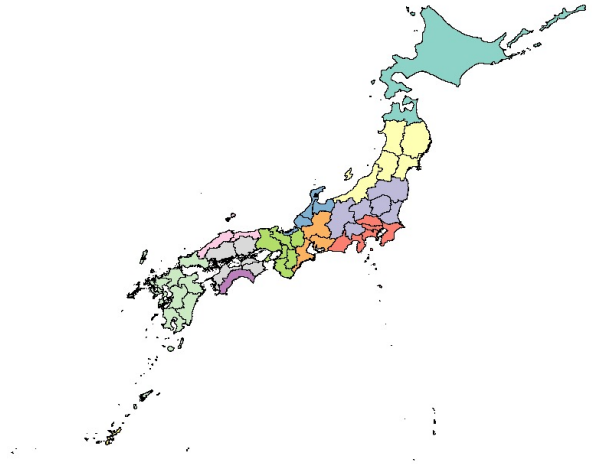
2016～2018年
３年平均



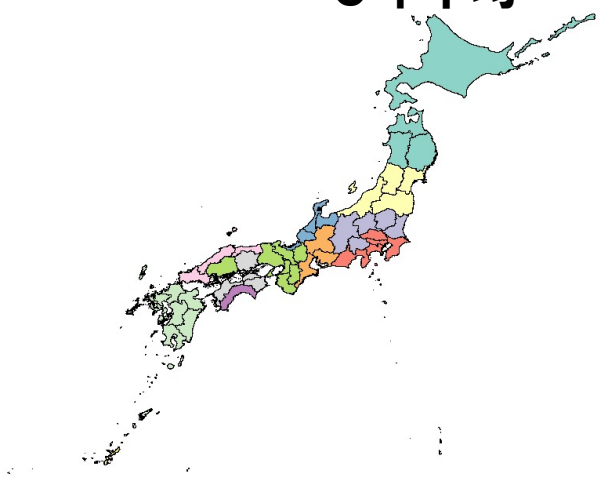
2015～2017年
３年平均



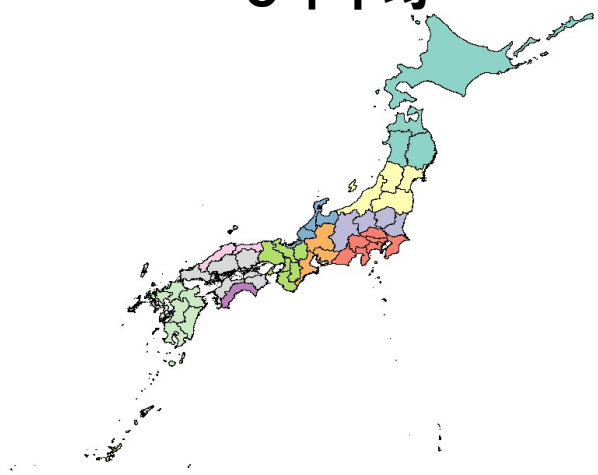
2012～2014年
３年平均



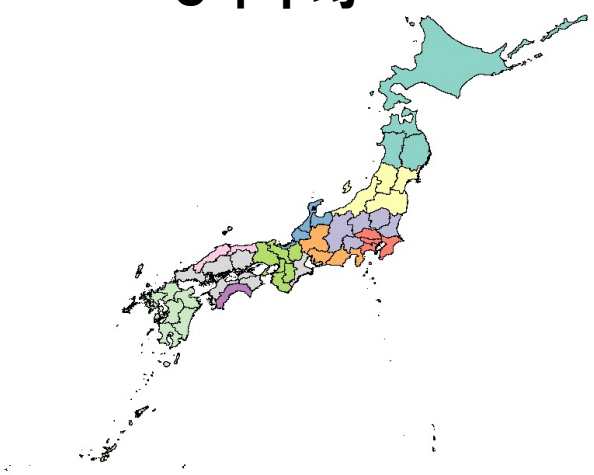
2011～2013年
３年平均



2010～2012年
３年平均



2007～2009年
３年平均



全てクラスター数を12で表示しました。

ほぼ同様な結果です。

分析結果の解釈が大切です

・家計調査データ(SSDSE-C-2020)のクラスター分析によって、日常的な食品嗜好の地域性が見事に見える化されました！（今回はward法を用いましたが、他の手法で分析するとどうなるでしょう？）

・この日本地域区分図をどう解釈してゆくか？　が最も大切な事です。

(解釈1)クラスター分割(凝集)において食料213品目がどう関与したのか？
→　「標準化データA」を眺めてみると傾向が見える
→スライド（参考１）

(解釈2)食品嗜好地域性を形作った背景（背後に流れる秘密）は何なのか？

家計調査データの数値解析結果だけを眺めていても(解釈2)はできません。

人文社会的な様々な知見を重ね合わせて考察する必要がある。
地理・歴史・食文化・民族移動・方言・交易・商流・
農林水産業・気候・商圏・物流・流通業界シェア
などなど
→スライド（参考２）

おわりに

今回ご紹介した分析事例は、
雑誌「エストレーラ」
2021年３月号に掲載されています

木村・高部(2021)

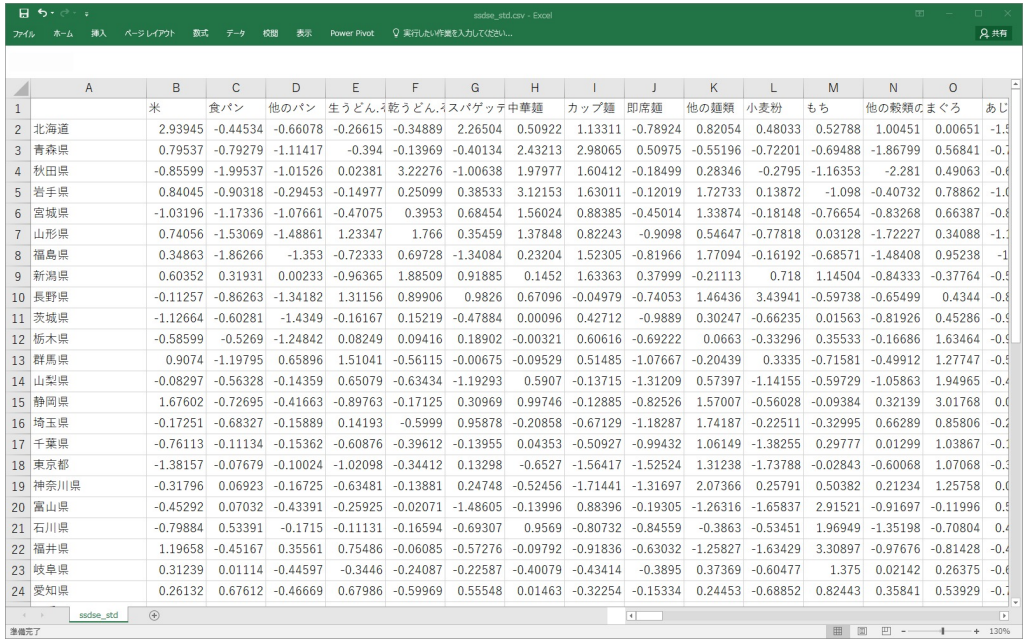
一連の図表や分析用Rプログラムについても
雑誌「エストレーラ」の発行元である
（公財）統計情報研究開発センター
のHPから自由にダウンロードが可能です。
是非、ご活用下さい。

https://www.sinfonica.or.jp/ 内の
［刊行物］ ＞ ［ESTRELA］ ＞ ［2.参考(掲載記事より)］ からダウンロード可能

木村・高部(2021):「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～」
『ESTRELA』324, pp.36-42, 統計情報研究開発センター

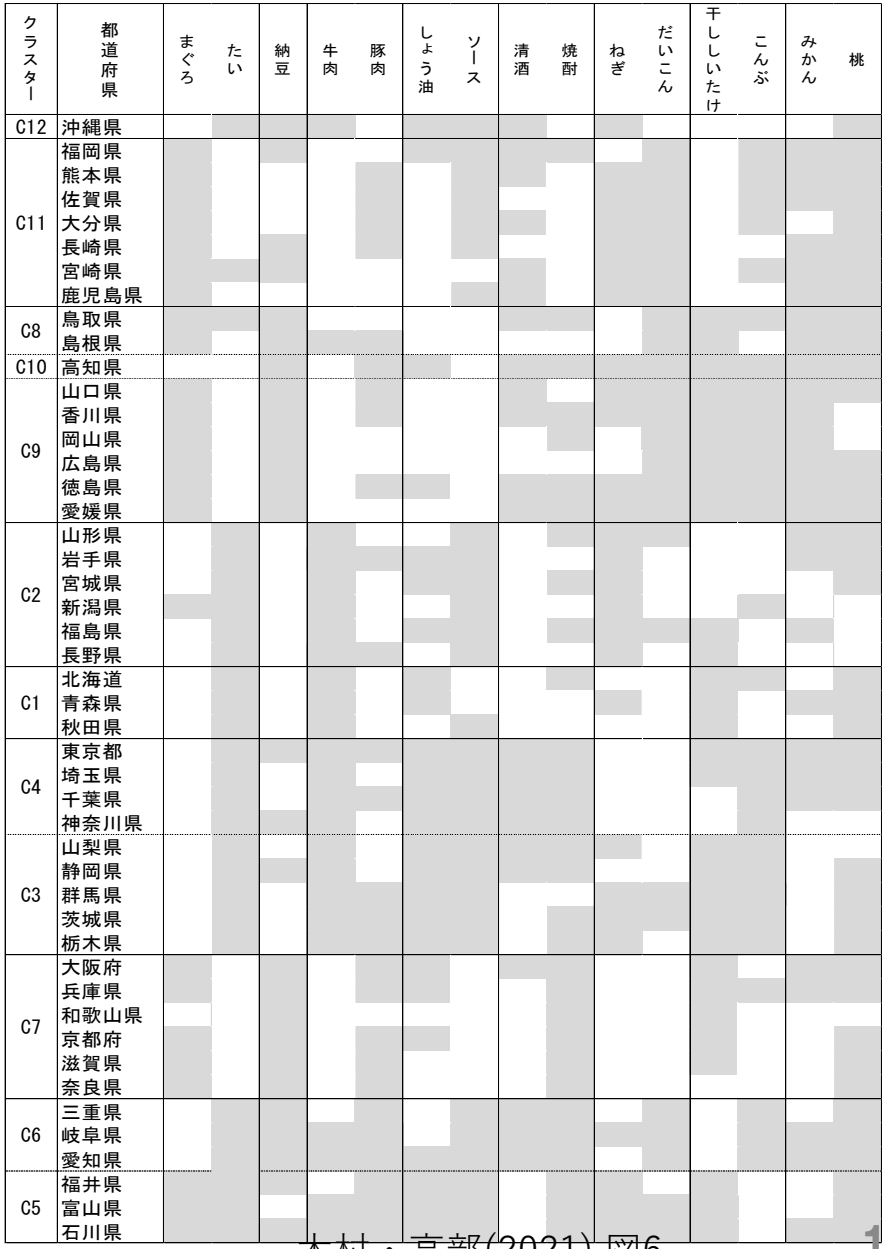
（参考1）「標準化データA」による品目関与の簡易的解釈

「標準化データA」



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		米	食パン	他のパン	生うどん,そ	乾うどん,そ	スパゲッテ	中華麺	カップ麺	即席麺	他の麺類	小麦粉	もち	他の穀類の	まぐろ
2	北海道	2.93945	-0.44534	-0.66078	-0.26615	-0.34889	2.26504	0.50922	1.13311	-0.78924	0.82054	0.48033	0.52788	1.00451	0.00651
3	青森県	0.79537	-0.79279	-1.11417	-0.394	-0.13969	-0.40134	2.43213	2.98065	0.50975	-0.55196	-0.72201	-0.69488	-1.86799	0.56841
4	秋田県	-0.85599	-1.99537	-1.01526	0.02381	3.22276	-1.00638	1.97977	1.60412	-0.18499	0.28346	-0.2795	-1.16353	-2.281	0.49063
5	岩手県	0.84045	-0.90318	-0.29453	-0.14977	0.25099	0.38533	3.12153	1.63011	-0.12019	1.72733	0.13872	-1.098	-0.40732	0.78862
6	宮城県	-1.03196	-1.17336	-1.07661	-0.47075	0.3953	0.68454	1.56024	0.88385	-0.45014	1.33874	-0.18148	-0.76654	-0.83268	0.66387
7	山形県	0.74056	-1.53069	-1.48861	1.23347	1.766	0.35459	1.37848	0.82243	-0.9098	0.54647	-0.77818	0.03128	-1.72227	0.34088
8	福島県	0.34863	-1.86266	-1.353	-0.72333	0.69728	-1.34084	0.23204	1.52305	-0.81966	1.77094	-0.16192	-0.68571	-1.48408	0.95238
9	新潟県	0.60352	0.31931	0.00233	-0.96365	1.88509	0.91885	0.1452	1.63363	0.37999	-0.21113	0.718	1.14504	-0.84333	-0.37764
10	長野県	-0.11257	-0.86263	-1.34182	1.31156	0.89906	0.9826	0.67096	-0.04979	-0.74053	1.46436	3.43941	-0.59738	-0.65499	0.4344
11	茨城県	-1.12664	-0.60281	-1.4349	-0.16167	0.15219	-0.47884	0.00096	0.42712	-0.9889	0.30247	-0.66235	0.01563	-0.81926	0.45286
12	栃木県	-0.58599	-0.5269	-1.24842	0.08249	0.09416	0.18902	-0.00321	0.60616	-0.69222	0.0663	-0.33296	0.35533	-0.16686	1.63464
13	群馬県	0.9074	-1.19795	0.65896	1.51041	-0.56115	-0.00675	-0.09529	0.51485	-1.07667	-0.20439	0.3335	-0.71581	-0.49912	1.27747
14	山梨県	-0.08297	-0.56328	-0.14359	0.65079	-0.63434	-1.19293	0.5907	-0.13715	-1.31209	0.57397	-1.14155	-0.59729	-1.05863	1.94965
15	静岡県	1.67602	-0.72695	-0.41663	-0.89763	-0.17125	0.30969	0.99746	-0.12885	-0.82526	1.57007	-0.56028	-0.09384	0.32139	3.01768
16	埼玉県	-0.17251	-0.68327	-0.15889	0.14193	-0.5999	0.95878	-0.20858	-0.67129	-1.18287	1.74187	-0.22511	-0.32995	0.66289	0.85806
17	千葉県	-0.76113	-0.11134	-0.15362	-0.60876	-0.39612	-0.13955	0.04353	-0.50927	-0.99432	1.06149	-1.38255	0.29777	0.01299	1.03867
18	東京都	-1.38157	-0.07679	-0.10024	-1.02098	-0.34412	0.13298	-0.6527	-1.56417	-1.52524	1.31238	-1.73788	-0.02843	-0.60068	1.07068
19	神奈川県	-0.31796	0.06923	-0.16725	-0.63481	-0.13881	0.24748	-0.52456	-1.71441	-1.31697	2.07366	0.25791	0.50382	0.21234	1.25758
20	富山県	-0.45292	0.07032	-0.43391	-0.25925	-0.02071	-1.48605	-0.13996	0.88396	-0.19305	-1.26316	-1.65837	2.91521	-0.91697	-0.11996
21	石川県	-0.79884	0.53391	-0.1715	-0.11131	-0.16594	-0.69307	0.9569	-0.80732	-0.84559	-0.3863	-0.53451	1.96949	-1.35198	-0.70804
22	福井県	1.19658	-0.45167	0.35561	0.75486	-0.06085	-0.57276	-0.09792	-0.91836	-0.63032	-1.25827	-1.63429	3.30897	-0.97676	-0.81428
23	岐阜県	0.31239	0.01114	-0.44597	-0.3446	-0.24087	-0.22587	-0.40079	-0.43414	-0.3895	0.37369	-0.60477	1.375	0.02142	0.26375
24	愛知県	0.26132	0.67612	-0.46669	0.67986	-0.59969	0.55548	0.01463	-0.32254	-0.15334	0.24453	-0.68852	0.82443	0.35841	0.53929

数値の符号で色分けすると、クラスター毎の共通品目が見えてくる。（2元ヒートマップならより詳細な情報が得られます。）



木村・高部(2021) 図6

(参考2) 糸口の例　方言区画図 (井上1986) との類似性

金田一春彦他編（1988）
『日本語百科大事典』大修館書店　より転載

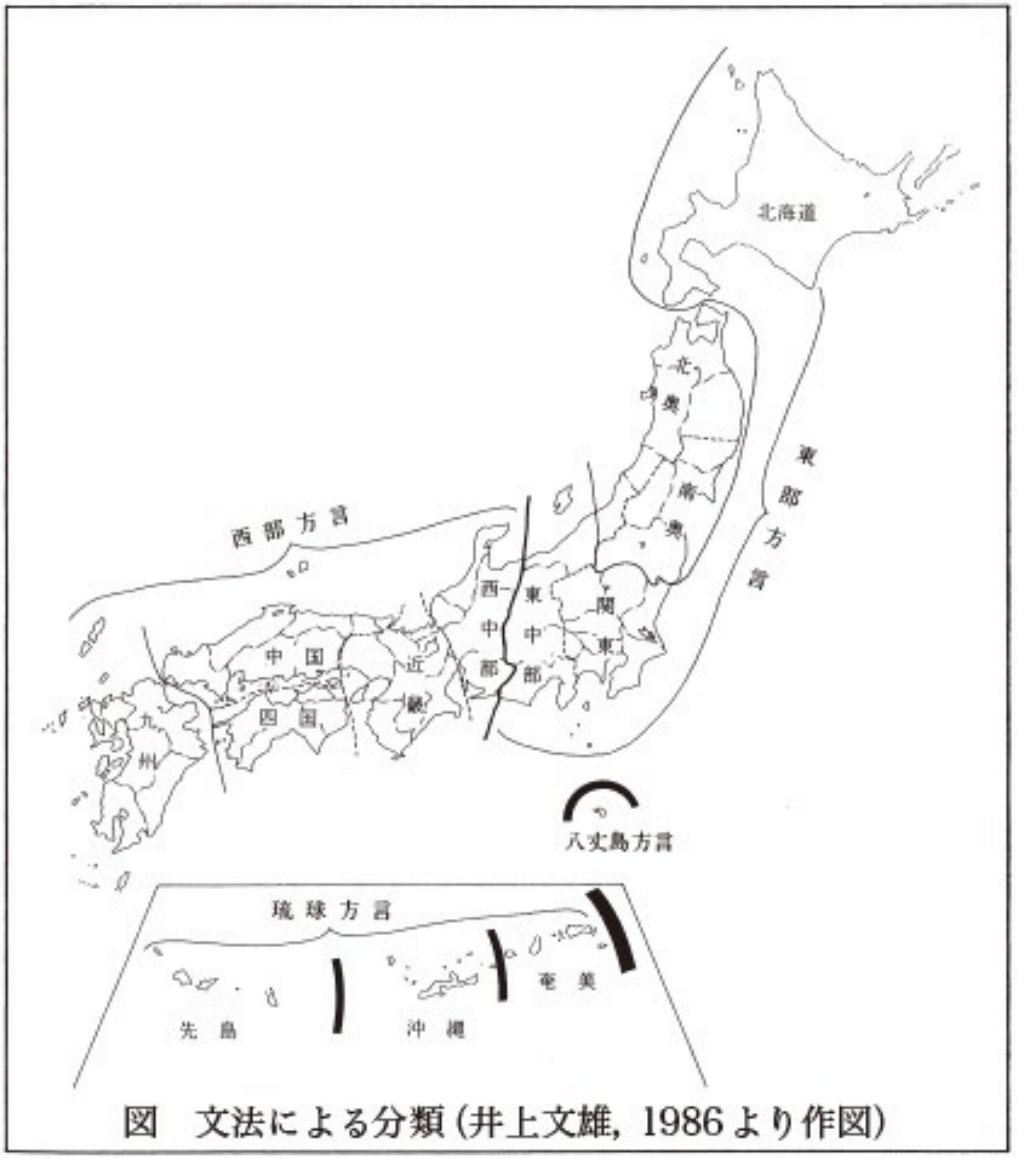


図　文法による分類 (井上文雄, 1986より作図)

木村・高部(2021)　図3

方言区分と食の地域性区分を比べてみると、海の果たす役割が全く異なっている？？？

なぜだろう？

(参考3) ウォード法の数式による表現

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

＜ward法でのクラスター凝集アルゴリズム＞

- あるクラスターiについて、その要素数をNi 、j番目の要素ベクトルを$\mathbf{x}_{ij}$ 、クラスターiの重心ベクトルを$\mathbf{c}_i$ 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、$\mathbf{x}_{ij}$ も$\mathbf{c}_i$ もm次元ベクトルである。

$$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijm}) \text{ 、 } \mathbf{c}_i = (c_{i1}, c_{i2}, \ldots, c_{im})$$

各要素がクラスターiに凝集した際の「情報の損失」(loss of information)の事をESSi(error sum of squares)で表現することにして、次の式で定義する。$\| \mathbf{x} \|$ は、ベクトル$\mathbf{x}$ の長さ（ユークリッド距離）とする。

$$\mathrm{ESSi} = \sum_{j=1}^{Ni} \|\mathbf{x}_{ij} - \mathbf{c}_i\|^2$$

- 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』が最も少ないpとqのペアを選定して凝集させる。つまり、

$$\min\{ \mathrm{ESSr} - (\mathrm{ESSp} + \mathrm{ESSq}) \}$$

なるpとqを凝集させてrとする。これをクラスターが１つになるまで繰り返す。　（以上）