

SASお客様向けwebinarシリーズ
機械学習によるビッグデータ分析の手法

#3 クラスター分析による分類（2） 階層的クラスタリング

2021年12月15日



Agenda

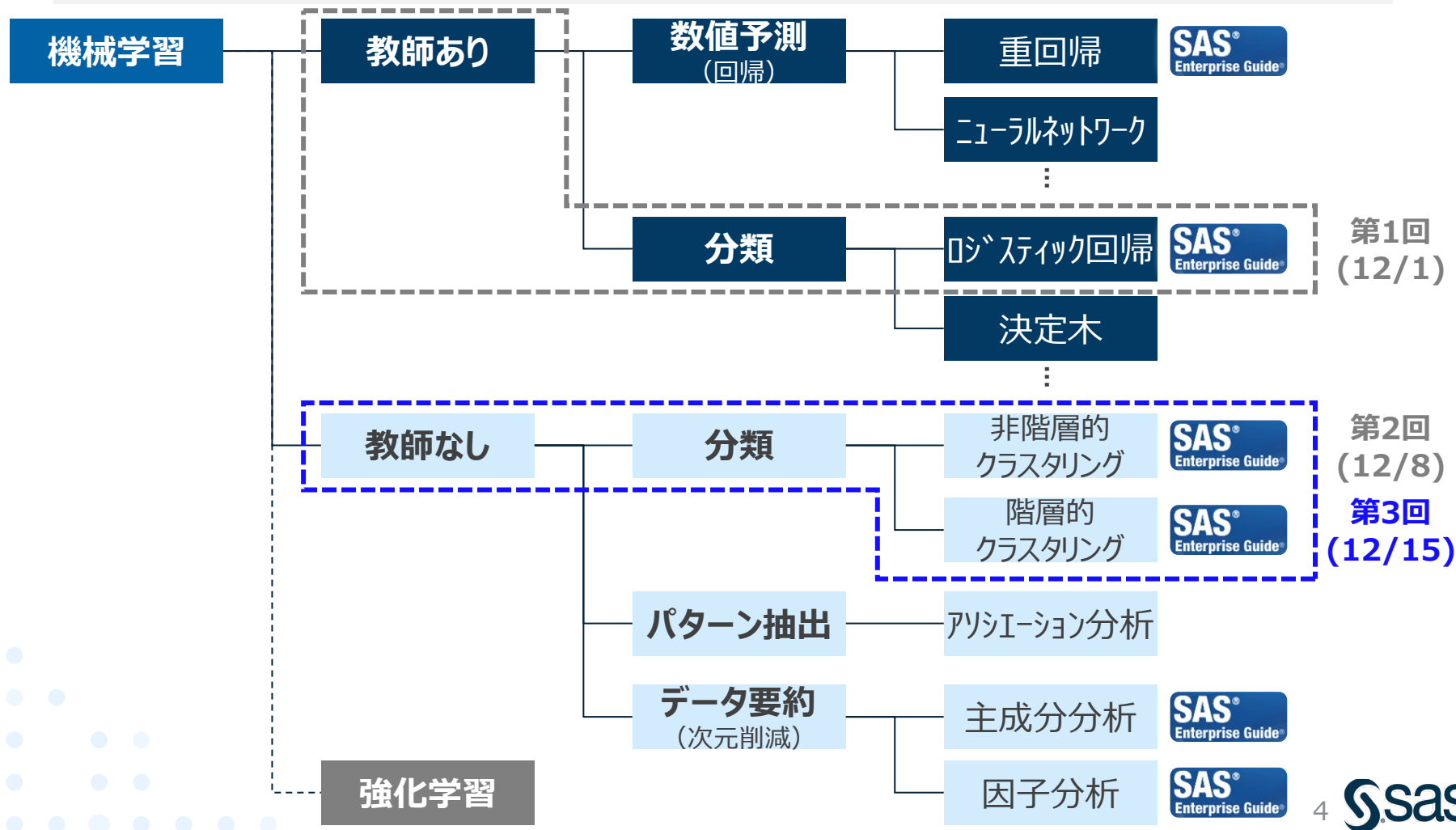
- クラスター分析による分類（２）：階層的クラスタリング
 - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
 - 各クラスタの解釈、予測への適用方法
 - 樹形図（デンドログラム）とクラスタ数の検討
 - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する
- 今後のデータサイエンス学習に向けて
 - データサイエンティストに求められるスキル
 - 学習リソースの紹介
 - SAS内サンプルデータの紹介と使い方

Agenda

- クラスター分析による分類（2）：階層的クラスタリング
 - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
 - 各クラスターの解釈、予測への適用方法
 - 樹形図（デンドログラム）とクラスター数の検討
 - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する
- 今後のデータサイエンス学習に向けて
 - データサイエンティストに求められるスキル
 - 学習リソースの紹介
 - SAS内サンプルデータの紹介と使い方

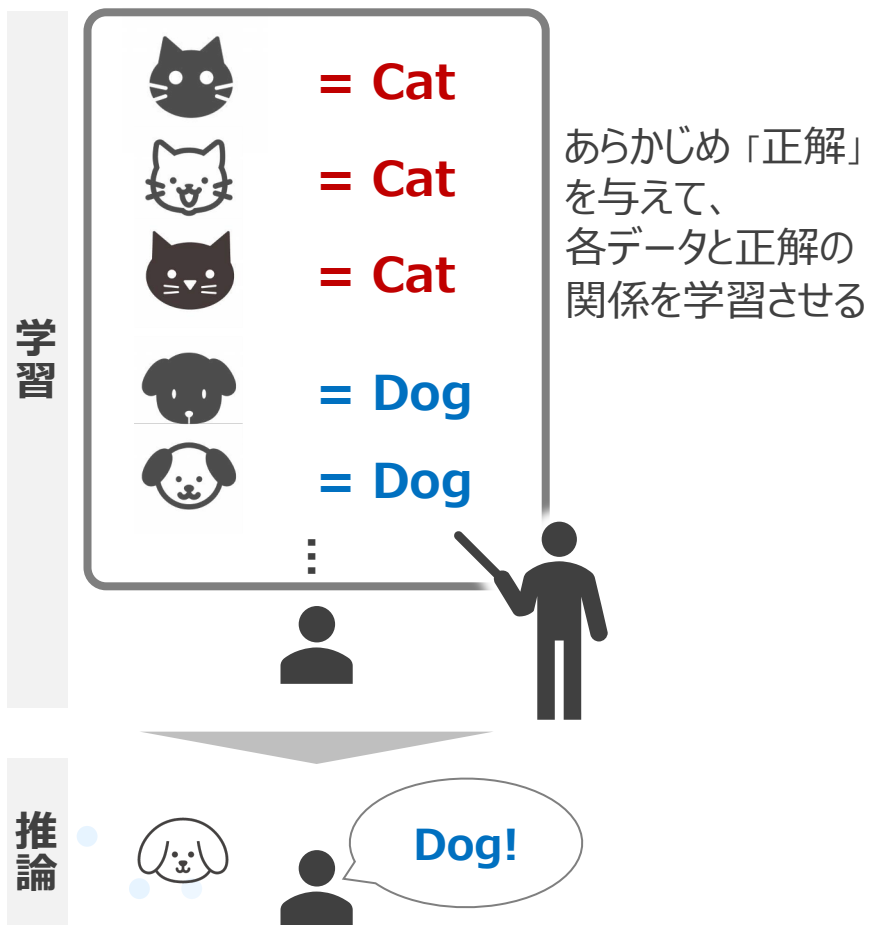
代表的な機械学習手法

- 機械学習手法は、教師あり、教師なし、強化学習に大別される
- なかでも、**教師あり分類**、**教師なし分類**は極めて基本的かつ頻用される手法である



教師あり学習 と 教師なし学習

教師あり学習



教師なし学習



教師なし学習のイメージ (クラスタリング)

- 各データ間の距離に基づき、近接データ (= 類似度が高いデータ) 同士のグループ (クラスタ) を作り、データを分類する手法
- 学習データなし**でデータを大きく層別したい場合に有効

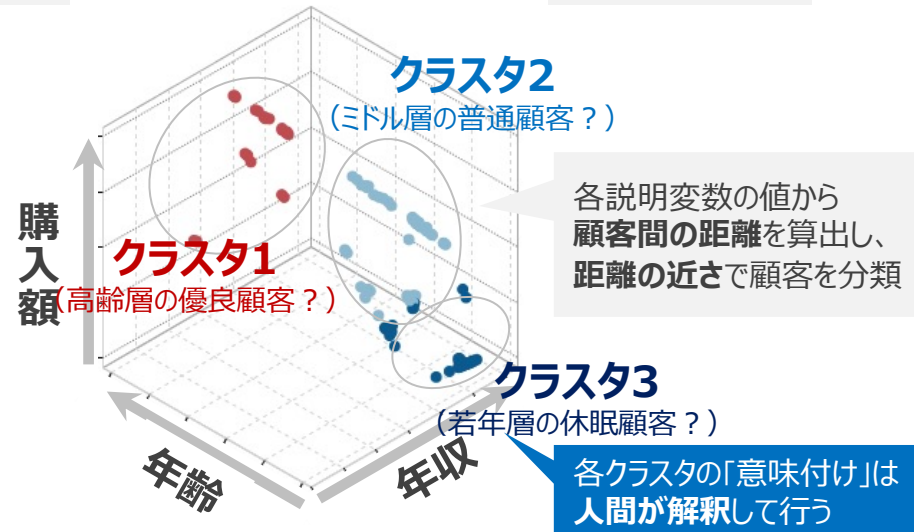
データ例

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...

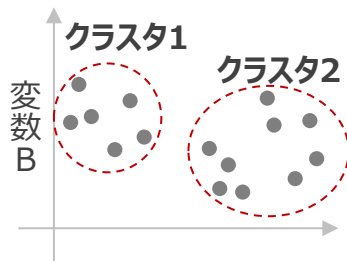
説明変数

※目的変数は無し

クラスタリング



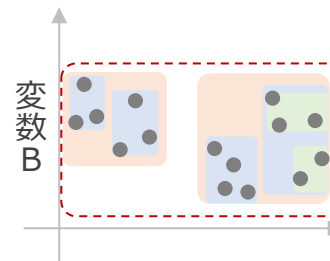
非階層的クラスタリング



主な手法

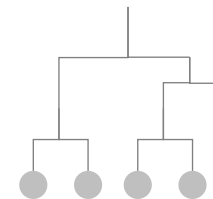
- **k-means法**
(k平均法)
- 混合ガウス

階層的クラスタリング



主な手法

- 最短距離法
- 最長距離法
- 群平均法
- ウォード法



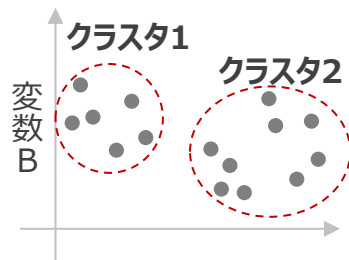
クラスタリング手法の種類



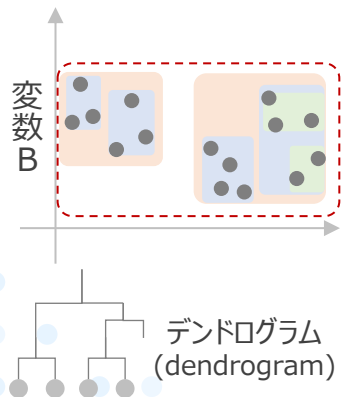
- クラスタリング手法は、「**非階層的**」と「**階層的**」に大別される
- 階層的クラスタリングはさらに **凝集型** と **分割型** があり、凝集型が用いられるのが一般的

手法の分類

非階層的クラスタリング



階層的クラスタリング



手法

▪ k-means法 (k平均法)

クラスタ内データの平均値をクラスタ重心として、距離に基づき、事前に設定したクラスタ数k個に分割



▪ その他

混合ガウス法、超体積法など

第2回で説明

似ている (≒距離の近い) データ／クラスタ同士を逐次まとめる (ボトムアップアプローチ)

▪ ウォード法

クラスタ内のデータの平方和を最小にするように併合



▪ 最短距離法 (最近隣法)

距離の近いデータから順番に併合

▪ 最長距離法 (最遠隣法)

距離の遠いデータから順番に併合

本日
ご説明

▪ 重心法

クラスタ重心からの距離に基づき併合



▪ 群平均法

各クラスタ同士で全データの距離の平均を基準に併合



▪ その他

メディアン法、可変法

凝集型

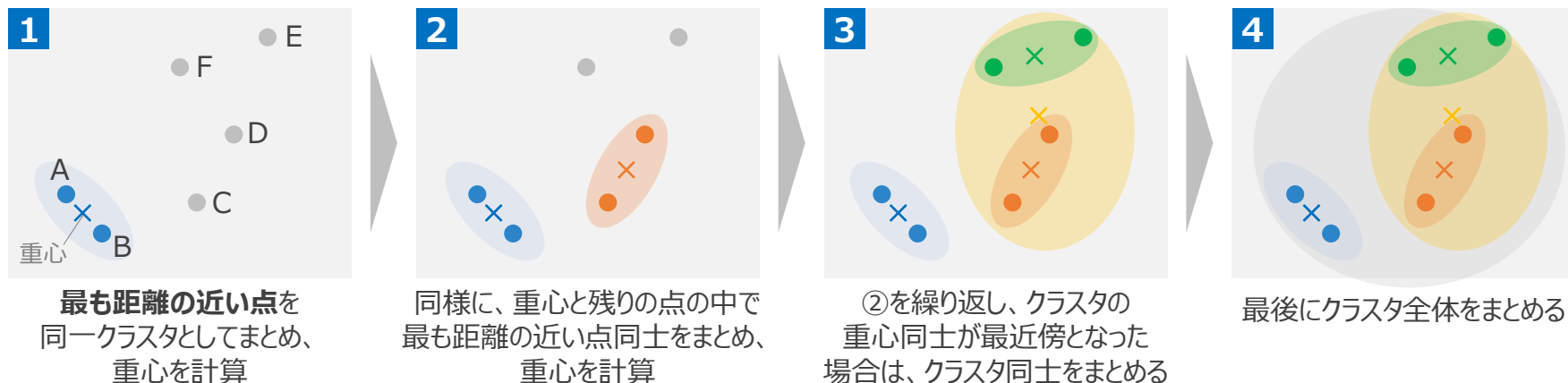
分割型

似ていないデータ／クラスタ同士を逐次分離させる (トップダウンアプローチ)

▪ Diana法

代表的な階層的クラスタリング： 凝集型階層クラスタリング

- 凝集型階層クラスタリングは、距離に応じて小さいクラスタを束ねて階層的に分類する手法
- クラスタ数は自動的に決定してくれる他、分類過程を可視化した**樹形図**（デンドログラム）も同時に出力されるので、結果の解釈やクラスタ数の決定に役立つ



凝集型階層的クラスタリング

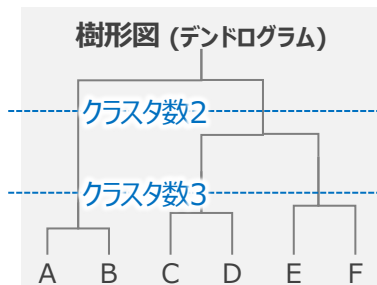
(agglomerative hierarchical clustering)

メリット

- クラスタ数は自動決定
- 樹形図により**分類過程が可視化**されることで、妥当なクラスタ数を人が判断可能

デメリット

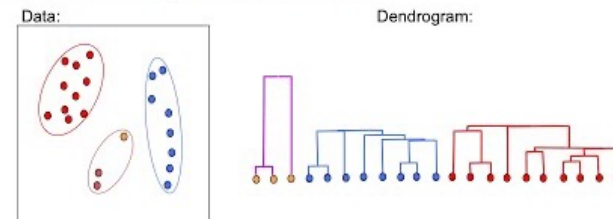
- 計算量が膨大
- データ量が多い場合、樹形図が複雑となり、解釈が困難になる



Copyright © SAS Institute Inc. All rights reserved.

Iteration m-3

Builds up a sequence of clusters ("hierarchical")



In matlab: "linkage" function (stats toolbox)

Algorithmic Complexity: $O(m^2 \log m) + (m-3) \cdot O(m \log m) +$

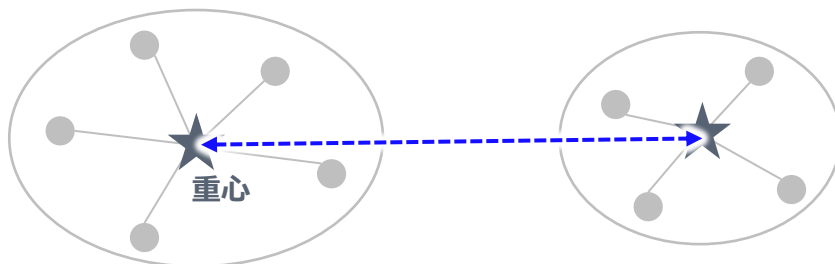
出典: <https://youtu.be/OcoE7JlbXvY>

「近い」の評価尺度バリエーション

- クラスタ間の「近さ」を測る指標には様々あるが、一概にどれが良いとは言えないため、**複数試して比較**するのが一般的である。ただし、一般には、群平均法やWard法（次頁）が頻用される
- 最短距離／最長距離法は、計算量が少なくて済む反面、1点の影響を大きく受けやすい

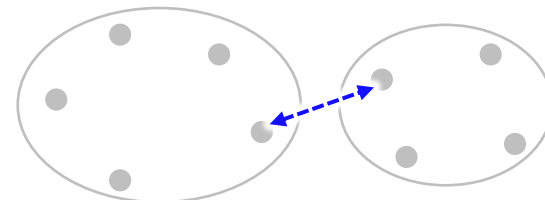
重心法

重心間の距離が近いクラスタを結合



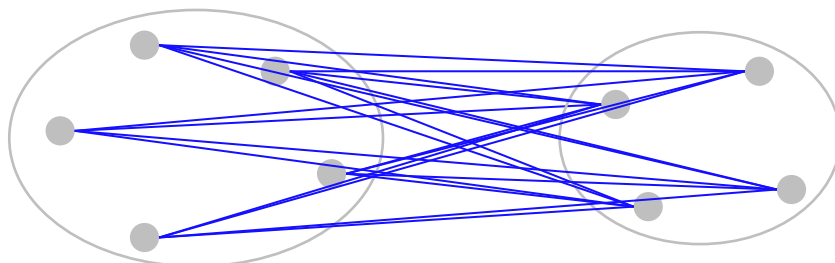
最短距離法

2つのクラスタ間で**最近傍**のデータをクラスタ間距離として採用



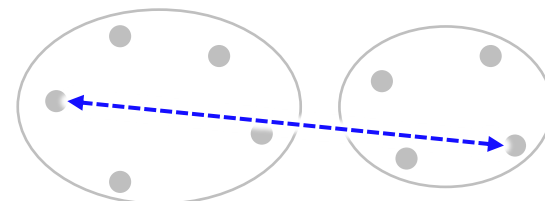
群平均法

クラスタ間で全データ間の距離を算出し、その**平均値**が近いクラスタを結合



最長距離法

2つのクラスタ間で**最遠方**のデータをクラスタ間距離として採用



Ward法の考え方

- Ward法*は最もよく用いられる手法であり、計算量が多いが、各データ点とクラスタ重心との関係性まで評価しているため、他手法に比べ、**分類感度が高い**とされる

*米国の統計学者Joe H. Ward, Jr.が1963年に発表した論文にちなむ

1

「クラスタ重心」と、「当該クラスタ内の各データ」との距離の総和（二乗和）をクラスタごとに算出

クラスタAの場合

$$A = a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2$$

クラスタBの場合

$$B = b_1^2 + b_2^2 + b_3^2 + b_4^2$$

A, Bの結合を仮定した場合のクラスタAB

2

注目する2つのクラスタを結合した場合を仮定し、「結合後のクラスタ重心」と「当該クラスタ内の各データ」との距離の総和（二乗和）を算出

$$AB = a_1'^2 + a_2'^2 + a_3'^2 + a_4'^2 + a_5'^2 + b_1'^2 + b_2'^2 + b_3'^2 + b_4'^2$$

3

1と2の差、つまり、 **$AB - (A+B)$** が**最小**となるクラスター結合を採用（結合前後でクラスタ内のばらつきに変化なし→統合してもOKと判定）

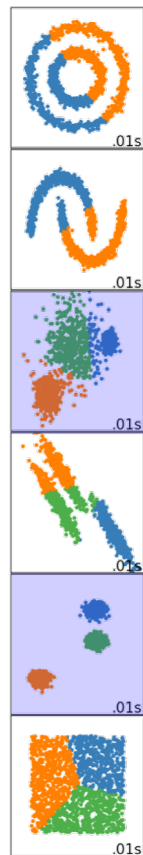
※近くにあり、ばらつきの小さいクラスタ同士が結合しやすい

参考：クラスタリング手法における分類結果の比較

- クラスタリング手法によって得意なデータパターンは異なり、様々な手法を試しながら、最適な手法を選択することが望ましい。中でも、k-meansは「重心からの距離」を用いて分類するため、円状のデータには強いが、楕円状や曲線状のデータは苦手

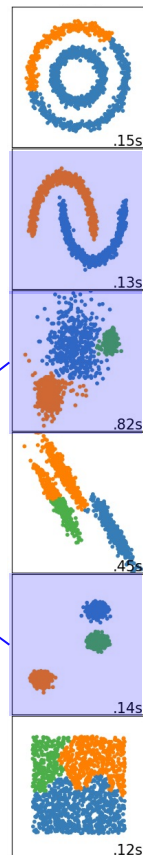
入力データ形式による違い

非階層的クラスタリング
(k-means)

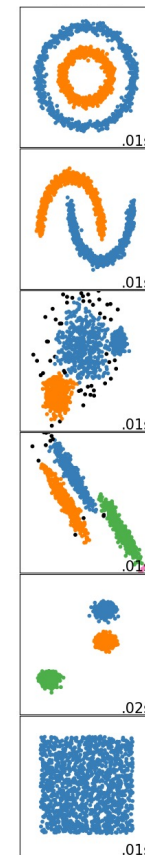


円状のデータには強い

階層的クラスタリング
(Ward法)



その他の参考手法：
DBSCAN



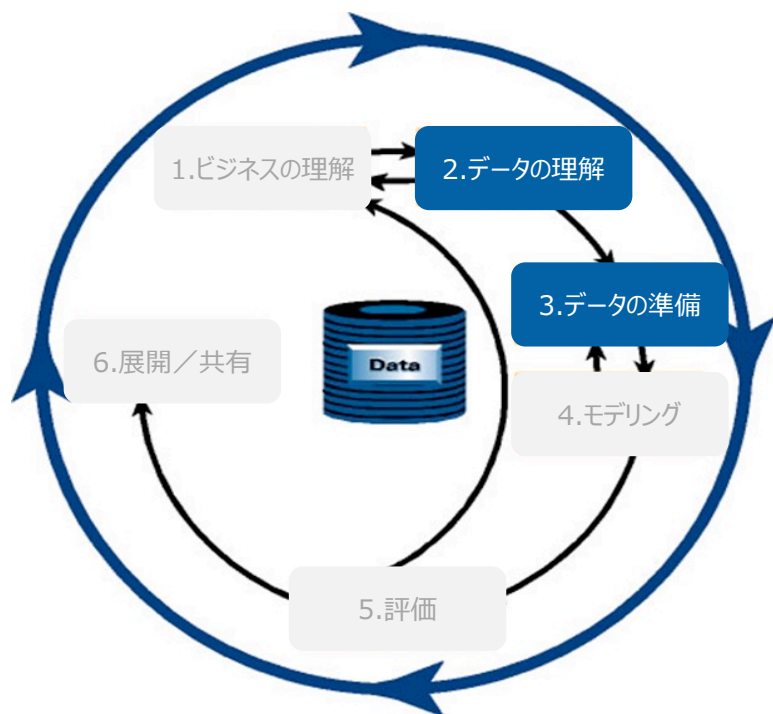
SAS Enterprise Guide での実装方法

- データの読み込み
- 階層的クラスタリング（Ward法、重心法、群平均法）
- 標準化したクラスタリング

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(CRoss Industry Standard Process for Data Mining)

1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル作成

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

使用データ（e-Statについて）

- 政府が公開する政府統計のオープンデータ “e-Stat” のデータを活用する
- 今回扱うデータの他にも、様々な統計データが公開されているので、企業内のデータと組み合わせることで、さらなる付加価値を生む可能性がある

The screenshot displays the e-Stat website, a portal for Japanese government statistics. At the top, the e-Stat logo is accompanied by the tagline '統計で見る日本' (Japan seen through statistics) and a description: 'e-Statは、日本の統計が閲覧できる政府統計ポータルサイトです' (e-Stat is a government statistics portal site where Japanese statistics can be browsed). Navigation links for 'お問い合わせ' (Contact), 'ヘルプ' (Help), and 'English' are in the top right, along with 'ログイン' (Login) and '新規登録' (New Registration) buttons. A blue navigation bar contains links: '統計データを探す' (Find statistics data), '統計データの活用' (Use statistics data), '統計データの高度利用' (Advanced use of statistics data), '統計関連情報' (Statistics-related information), and 'リンク集' (Link collection).

The main content area is divided into two primary sections:

- 統計データを探す (Find statistics data)** (政府統計の調査結果を探します): This section includes three filters: 'すべて' (All) for browsing all government statistics, '分野' (Field) for searching by 17 statistical fields, and '組織' (Organization) for searching by government agencies. A search bar with the example '例：国勢調査' (Example: Census) and a '検索' (Search) button is provided. A 'その他の絞り込み' (Other filters) button is also present.
- 統計データを活用する (Use statistics data)**: This section offers four ways to utilize data: 'グラフ' (Graph) for displaying main indicators on a dashboard, '時系列表' (Time series table) for displaying main indicators in a table, '地図' (Map) for displaying data on a map (Statistics GIS), and '地域' (Region) for displaying main data for prefectures, cities, and towns.

On the right side, a vertical menu provides additional resources: '利用ガイド' (User Guide), '統計データの高度利用' (Advanced use of statistics data) with a sub-link for 'マイクロデータの利用' (Use of microdata), '開発者向け' (For developers) for API and LOD data, and '統計関連情報' (Statistics-related information) with a link for '統計分類・調査計画等' (Statistics classification, survey plans, etc.).

Source: <https://www.e-stat.go.jp/>

使用データ

- 今回は、このうち、5年に1度実施している「全国消費実態調査」（現在の名称は「全国家計構造調査」）のデータを用いて、**都道府県別の消費動向から、類似の都道府県をグルーピングすること**を考える

政府統計名	全国家計構造調査（旧全国消費実態調査）	詳細		
提供統計名	平成26年全国消費実態調査			
提供分類1	全国			
提供分類2	家計収支に関する結果			
提供分類3	総世帯			

表番号	統計表	調査年月	公開（更新）日	表示・ダウンロード
フロー編				
42	年間収入階級・年間収入十分位階級別 1世帯当たり 1か月間の収入と支出			
	総世帯	2014年	2015-12-16	EXCEL DB
	勤労者世帯	2014年	2015-12-16	EXCEL DB
43	世帯主の年齢階級別 1世帯当たり 1か月間の収入と支出			
	総世帯・勤労者世帯	2014年	2015-12-16	EXCEL DB
44	住居の所有関係別 1世帯当たり 1か月間の収入と支出			
	総世帯・勤労者世帯	2014年	2015-12-16	EXCEL DB
45	資産の種類・資産額階級別 1世帯当たり 1か月間の収入と支出（純資産）			
	総世帯	2014年	2016-03-25	EXCEL DB
	勤労者世帯	2014年	2016-03-25	EXCEL DB
	資産の種類・資産額階級別 1世帯当たり 1か月間の収入と支出（総資産）			
	総世帯	2014年	2016-03-25	EXCEL DB
	勤労者世帯	2014年	2016-03-25	EXCEL DB
地域編				
13	地域別 1世帯当たり 1か月間の収入と支出			
	総世帯	2014年	2015-12-16	EXCEL DB
	勤労者世帯	2014年	2015-12-16	EXCEL DB

Source: <https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200564&tstat=000001073908&cycle=0&tclass1=000001073965&tclass2=000001074840&tclass3=000001077457&tclass4val=0>

データの概要 (加工前)

- e-Statより素データをダウンロードして開くと、開始行や開始列がずれていたり、空白行があったりと、加工が必要な形式であることがわかる
- 今回は、本データから都道府県別の消費細目データ部分を抽出し、加工済のデータを用いる

消費の細目

第 13 表 地域別1																
Table 13. Monthly Receipts and Dis																
都道府県データ																
Total Households																
都道府県 Prefectures																
都道府県 (続き)																
北海道	青森県	岩手県	宮城県	秋田県	山形県	福島県	茨城県	栃木県	群馬県	埼玉県	千葉県	東京都	神奈川県	新潟県		
Hokkaido	Aomori	Iwate	Miyagi	Akita	Yamagata	Fukushima	Ibaraki	Tochigi	Gumma	Saitama	Chiba	Tokyo	Kanagawa	Niigata		
集計世帯数	2,164	739	746	789	770	757	877	1,424	877	905	2,878	2,759	2,563	2,671		
世帯数分布(抽出率調整)	2,374,215	476,970	507,292	923,420	375,617	379,431	757,313	1,076,891	722,506	727,684	2,850,919	2,467,193	6,157,574	3,918,553	849,711	
(1万の割合)	459	92	98	178	73	73	146	208	140	141	551	477	1,190	757		
世帯主の年齢(歳)	2.21	2.52	2.54	2.49	2.53	2.73	2.59	2.63	2.57	2.63	2.49	2.43	2.14	2.29		
18歳未満の世帯主(人)	0.36	0.40	0.37	0.38	0.35	0.40	0.43	0.43	0.42	0.44	0.42	0.39	0.35	0.35		
65歳以上の世帯主(人)	0.64	0.76	0.82	0.77	0.80	0.83	0.77	0.68	0.72	0.71	0.65	0.69	0.58	0.61		
うち無職の世帯主(人)	0.54	0.58	0.60	0.65	0.64	0.65	0.64	0.53	0.56	0.55	0.52	0.55	0.42	0.49		
有業の世帯主の年齢(歳)	0.97	1.25	1.29	1.12	1.29	1.44	1.22	1.29	1.23	1.30	1.17	1.11	1.07	1.09		
世帯主の性別(男)	0.721	0.752	0.744	0.727	0.789	0.801	0.785	0.843	0.820	0.786	0.795	0.808	0.742	0.784		
世帯主の性別(女)	0.279	0.248	0.256	0.273	0.211	0.199	0.215	0.157	0.180	0.214	0.205	0.192	0.258	0.216		
持ち家率(現住居)	75.7	77.2	80.8	73.2	80.6	80.8	75.9	83.6	83.9	82.5	77.4	80.9	67.6	71.7		
うち住宅ローン保有率(%)	22.4	16.1	20.1	19.3	16.9	22.0	17.1	21.9	23.9	23.2	25.3	24.8	23.3	23.5		
家賃・地代を支払っている世帯の割合(%)	25.5	20.9	20.7	24.0	19.8	18.4	21.2	16.4	18.1	15.8	21.7	19.6	34.3	26.4		
現住居の延べ床面積(m ²)	107.1	126.9	134.2	117.6	141.7	148.9	124.6	120.5	125.9	123.9	96.7	98.2	77.3	96.7		
自動車保有台数(千世帯あたり)	1,074	1,390	1,459	1,326	1,544	1,751	1,603	1,697	1,666	1,697	996	1,020	490	731		
自動車保有率(%)	77.5	83.4	82.0	82.0	88.8	92.5	90.9	90.8	92.9	91.7	73.0	74.9	43.6	61.8		
年間収入(千円)	4,592	4,451	4,993	4,994	5,153	5,821	5,085	5,676	5,576	5,449	5,678	5,752	6,004	5,762		
消費支出	231,757	215,712	239,630	251,694	229,200	266,205	245,341	269,082	275,745	255,178	265,501	269,003	278,284	269,216		
食料	54,281	55,180	57,514	59,052	59,146	63,042	57,891	62,433	63,866	61,858	64,632	66,536	68,380	67,197		
穀類	5,147	4,726	5,013	5,048	5,170	5,726	5,549	5,465	5,516	5,901	5,512	5,393	4,899	5,244		
米	2,113	1,742	2,004	2,090	2,550	2,443	2,528	2,237	2,146	2,359	1,814	1,796	1,369	1,575		
パン	1,760	1,633	1,632	1,296	1,627	1,619	1,781	1,779	2,000	2,130	2,175	2,128	2,197	1,171		
麺類	1,017	1,163	1,111	1,114	1,173	1,352	1,167	1,168	1,277	1,270	1,253	1,113	1,090	1,161		
他の穀類	258	204	264	213	152	304	235	280	315	272	315	309	313	310		
魚介類	5,137	5,543	5,651	5,223	5,828	5,390	5,047	5,255	5,286	5,239	4,979	5,401	4,886	5,123		
生鮮魚介類	3,073	3,379	3,451	3,451	3,451	3,451	3,451	3,451	3,451	3,451	3,451	3,451	3,451	3,451		

消費の細目に関するデータを活用

データの概要（加工後）

- e-Statより素データをダウンロードして開くと、開始行や開始列がずれていたり、空白行があったりと、加工が必要な形式であることがわかる
- 今回は、本データから都道府県別の消費細目データ部分を抽出し、加工済のデータを用いる

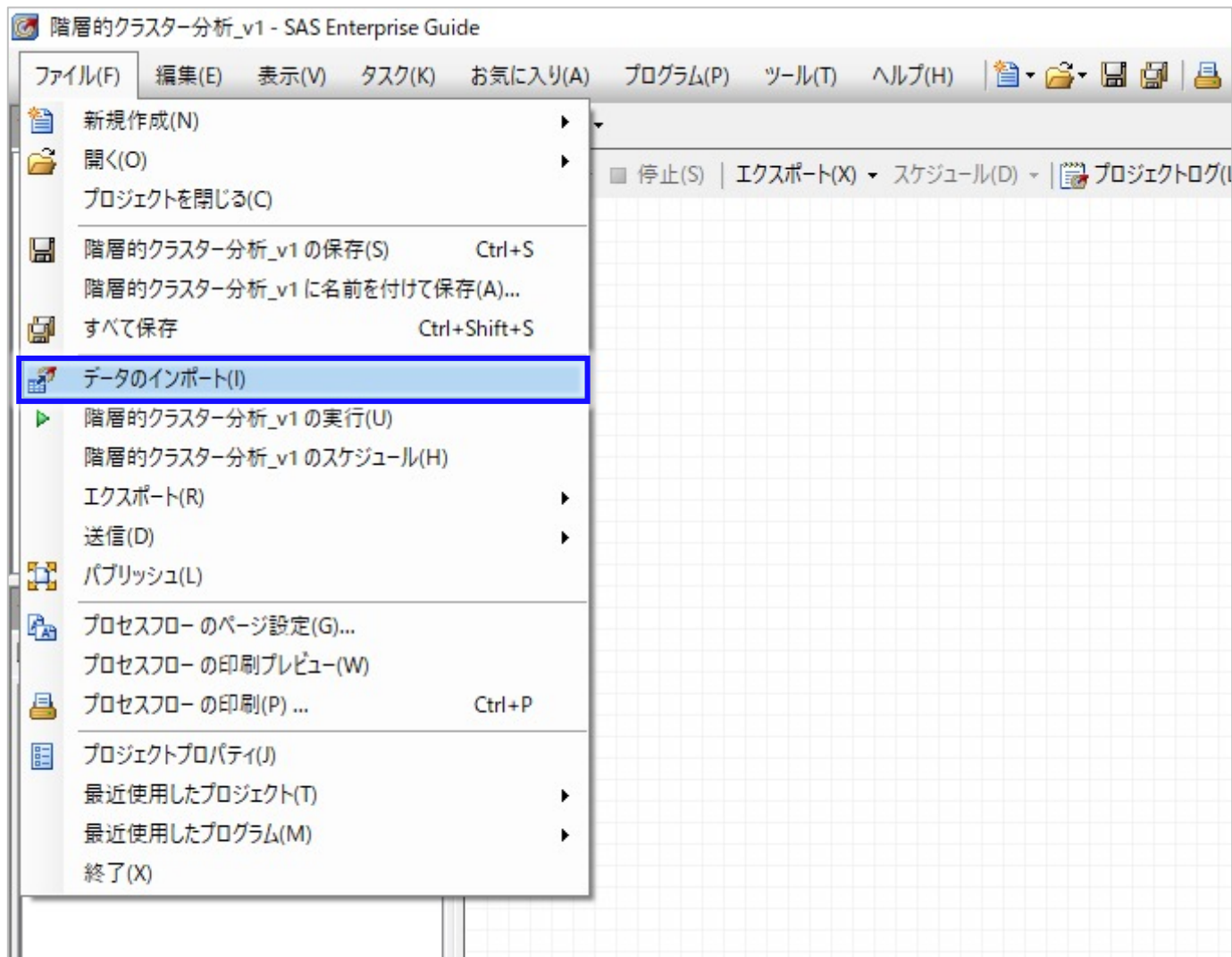
予測(分析)対象を説明するための変数

#	都道府県	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	読書	娯視・観覧	旅行	スポーツ	月謝	会費・つきあい費
1	北海道	54281	17491	19520	9089	10208	9234	35627	5630	23323	3894	2106	6506	2492	1430	1011
2	青森県	55180	14357	22420	9162	8972	9936	33039	4880	16564	3256	2084	3839	1153	777	1079
3	岩手県	57514	14782	21267	8973	8288	10273	42912	5805	20278	3534	2428	6326	1447	1219	1458
4	宮城県	59052	16353	20331	9700	10640	10851	40742	7331	23394	3241	2322	7734	2172	1587	1168
5	秋田県	59146	12875	22394	9108	7467	9214	37645	4472	21037	3349	2403	6672	1617	979	1191
6	山形県	63042	13186	24030	11771	9486	10407	40365	6667	24577	3706	2227	5438	1659	1333	2230
7	福島県	57891	12579	20102	9633	8949	9613	44195	4614	21216	3637	1921	6701	1743	1432	1027
8	茨城県	62433	17292	20330	9186	10645	10668	44764	10113	26592	3638	2451	7254	2639	2041	1134
9	栃木県	63866	18994	19997	9914	10252	12956	47208	6889	27561	3564	2723	7996	2908	2400	996
10	群馬県	61858	15629	18305	9605	9682	11270	43782	7539	25659	3767	2473	6800	2837	1965	931
11	埼玉県	64632	20131	17747	8544	11403	11616	40152	12953	29055	3800	3115	8887	2699	2705	632
12	千葉県	66536	17887	18039	8820	11826	11859	39048	12165	30385	4186	3512	10407	2889	2800	684
13	東京都	68380	33295	16315	8691	12404	12151	33118	11060	32038	4180	3882	14361	3061	2964	808
14	神奈川県	67197	22708	16957	8783	11591	11443	38440	11004	31833	4159	3607	10790	2874	2894	889
15	新潟県	64400	15713	21881	8900	9077	10628	38983	6736	23878	3605	2535	7327	1955	1715	1032
16	富山県	67635	14518	21894	10624	9387	10776	51532	7879	27246	4003	4101	5869	2139	2433	1373
17	石川県	66478	17678	18423	8733	9512	11135	42087	7993	27548	4220	3207	8516	1898	1966	1402
18	福井県	67429	12168	20741	9034	10204	11287	45576	9585	27984	3482	4031	8052	1863	2080	1539
19	山梨県	57641	17234	18209	7890	9429	10280	39392	9066	25849	3531	3699	6404	1877	2212	1132
20	長野県	62406	21145	21350	9866	9375	11987	42846	8047	27147	4035	2910	7265	2537	1830	1224
21	岐阜県	61939	12754	19952	9042	9942	10463	41580	7541	25777	3627	2610	6754	2280	2056	1306
22	静岡県	62396	15048	18407	9012	9985	11488	41580	7541	25777	3627	2610	6754	2280	2056	1306
23	愛知県	64248	21485	17573	9010	11051	11880	41580	7541	25777	3627	2610	6754	2280	2056	1306
24	三重県	63275	12856	19237	9036	11444	12889	36645	9066	26012	3984	3109	8260	1855	2174	985
25	滋賀県	63385	16479	18807	9587	10034	10818	36645	9066	26012	3984	3109	8260	1855	2174	985
26	京都府	65337	13829	17928	8409	12630	9239	31046	10348	25016	3744	3264	7492	2465	2250	658
27	大阪府	62386	18778	16292	7230	9898	10782	31046	10348	25016	3744	3264	7492	2465	2250	658
28	兵庫県	63620	19262	16725	8281	10712	10926	36040	9806	27000	3827	3023	8397	2546	2548	823
29	奈良県	66408	17630	19784	9875	11068	12405	42593	14481	27121	3849	3065	9467	2170	2609	986
30	和歌山県	58010	10696	17125	8152	9250	8326	36333	4901	23890	3376	2656	4995	2098	1787	888
31	鳥取県	58027	13626	18488	8143	9050	10320	41570	4966	24212	3198	3787	7077	1775	1865	904
32	島根県	59223	11926	19494	8915	8767	11814	40722	3866	23678	3446	3538	6335	1619	1512	1656
33	岡山県	58368	13776	18306	8286	9846	10347	38978	8451	24914	3052	2796	6682	2430	1789	931
34	広島県	58058	17721	17128	9180	9622	11195	38580	8773	24978	3308	2660	7944	1918	2004	997
35	山口県	55832	18576	16610	9381	8003	10961	35524	5193	23931	3557	2873	6234	1514	1798	840
36	徳島県	55896	16389	18015	8680	9656	10261	38507	6659	23923	3439	3762	7064	2039	1810	997
37	香川県	57352	15438	17319	8338	8754	11070	40876	6059	25565	3476	2814	6197	2689	2176	942
38	愛媛県	55531	13489	17201	8171	8284	9224	32679	7901	19353	2872	2450	5539	1732	1774	937
39	高知県	54971	14463	16479	7609	7510	10329	32613	5206	20184	3273	2447	5228	1722	1167	867
40	福岡県	54633	18999	16314	8029	9823	10405	36057	7360	24134	3073	2790	9478	2489	1809	784
41	佐賀県	57104	13214	17556	8682	9647	11281	40406	6975	24864	3354	2813	6326	2333	1889	1189
42	長崎県	51798	18624	16853	7291	7934	10115	34480	6345	19631	2687	2421	7528	1383	1517	1013
43	熊本県	55006	11286	16802	8254	10041	11155	34633	6967	21046	2889	2309	5544	1688	2016	682
44	大分県	53558	14707	15685	8558	10853	11677	36458	3243	22105	3021	3223	5354	2139	1327	1153
45	宮崎県	53347	15963	15828	8228	8386	9476	36294	6276	21314	2565	3061	6208	2410	1413	1210
46	鹿児島県	50294	14792	15496	7800	7857	10022	39992	5063	18721	2593	2002	5533	2053	1160	1488
47	沖縄県	48770	22616	17251	6750	5010	8088	28055	5169	16217	2500	1492	3913	1767	1700	970

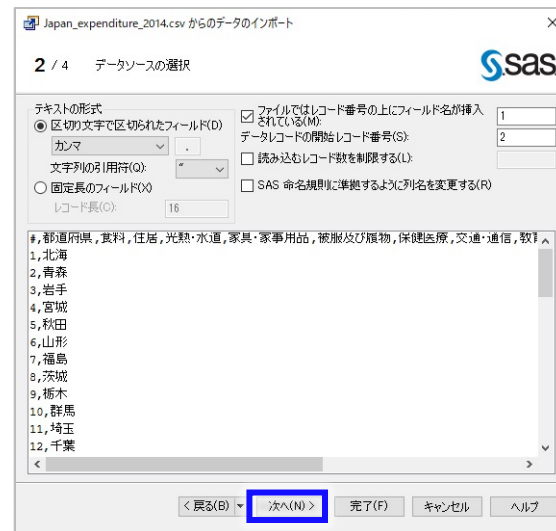
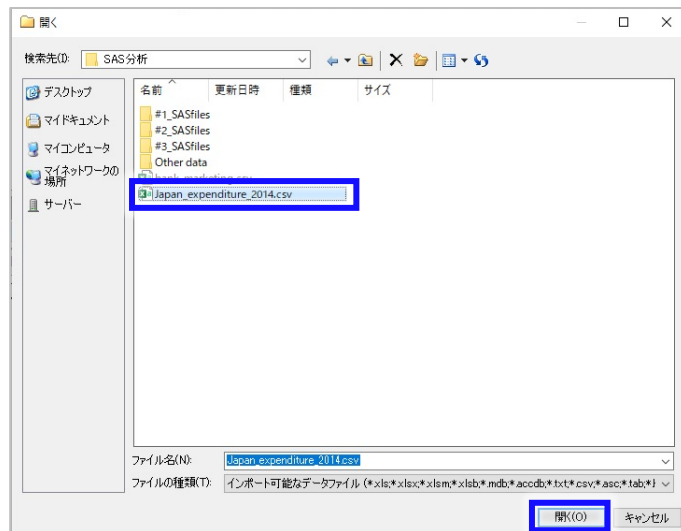
説明変数

都道府県

データの読み込み (1/2)

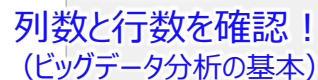
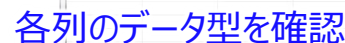


データの読み込み (2/2)





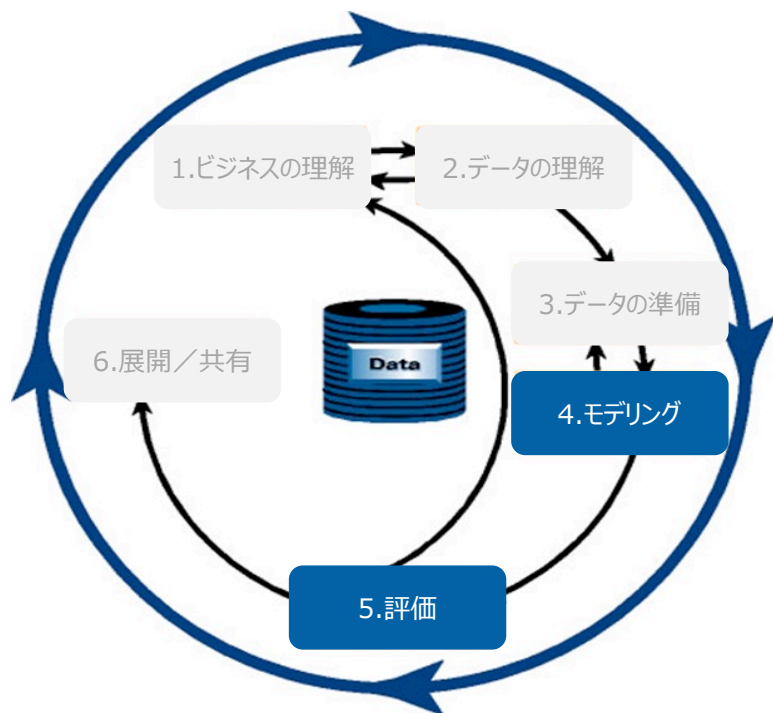
読み込まれたデータ



ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(CRoss Industry Standard Process for Data Mining)

1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル作成

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

階層的クラスタリング（群平均法）－ 実行方法（1/3）

[分析] → [多変量解析] → [クラスター分析] をクリック

The screenshot shows the SAS Enterprise Guide interface. The 'Analyze' menu is open, and the 'Multivariate' option is selected. The 'Cluster Analysis' option is highlighted in the 'Multivariate' submenu. The background shows a data table with columns for various categories like '食料' (Food), '住居' (Housing), '光熱・水道' (Utilities), '教育' (Education), etc., and rows of numerical data.

食料	住居	光熱・水道	教育	教養娯楽	読書	聴視・観覧	旅行	スポーツ
54281	17491	19520	5630	23323	3894	2106	6506	2492
55180	14357	22420			3256	2084	3839	1153
57514	14782	21267			3534	2428	6326	1447
59052	16353	20331			3241	2322	7734	2172
59146	12875	22394			3349	2403	6672	1617
63042	13186	24030			3706	2227	5438	1659
57891	12579	20102			3637	1921	6701	1743
62433	17292	20330			3638	2451	7254	2639
63866	18994	19997			3564	2723	7996	2908
61858	15629	18305			3767	2473	6800	2837
64632	20131	17747			3800	3115	8887	2699
66536	17887	18039			4186	3512	10407	2889
68380	33295	16315			4180	3882	14361	3061
67197	22708	16957			4159	3607	10790	2874

階層的クラスタリング（群平均法） – 実行方法（2/3）

- ・ 左側の [変数リスト] より、数値型の変数を選択し、右側の [分析変数] にドラッグ&ドロップ
- ・ 左側の [変数リスト] より、都道府県を選択し、右側の [ラベルの識別] にドラッグ&ドロップ

Local:WORK.JAPAN_EXPENDITURE_2014 のクラスター分析

データ
データソース: Local:WORK.JAPAN_EXPENDITURE_2014
タスクフィルタ: なし

変数リスト(A)

名前

都道府県

住居

光熱・水道

家具・家事用品

被服及び履物

保健医療

交通・通信

教育

教養娯楽

読書

聴視・観覧

旅行

スポーツ

月謝

会費・つきあい費

タスクの役割(T):

分析変数

食料

住居

光熱・水道

家具・家事用品

被服及び履物

保健医療

交通・通信

教育

教養娯楽

読書

聴視・観覧

旅行

スポーツ

月謝

会費・つきあい費

グループ分析

度数カウント (選択の上限: 1)

ラベルの識別 (選択の上限: 1)

都道府県

ラベルの識別に
都道府県を設定
(出力結果を見やすくするため)

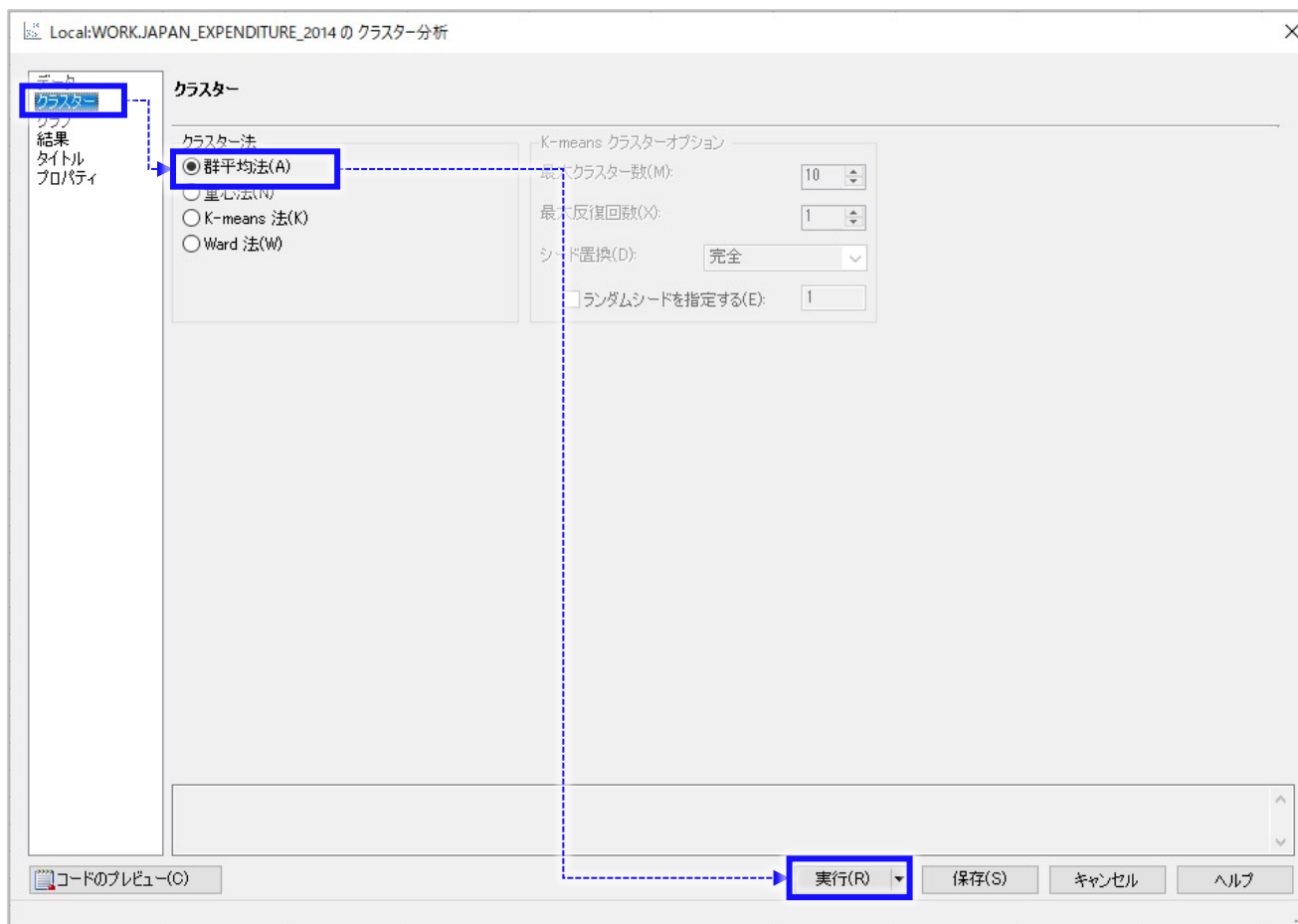
ドラッグ & ドロップ

コードのプレビュー(C)

実行(R) 保存(S) キャンセル ヘルプ

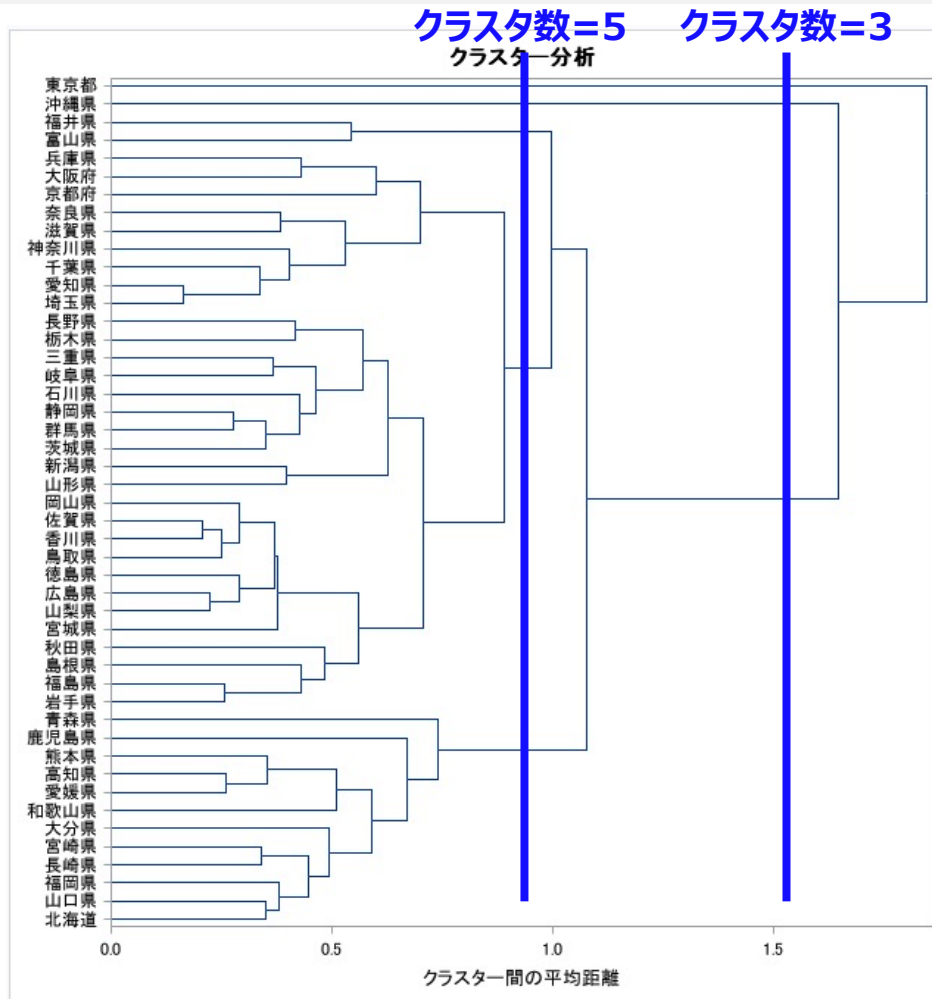
階層的クラスタリング（群平均法） – 実行方法（3/3）

- 左パネルから [クラスター] メニュー をクリック
- クラスタ法から [群平均法] を選択し、実行ボタン

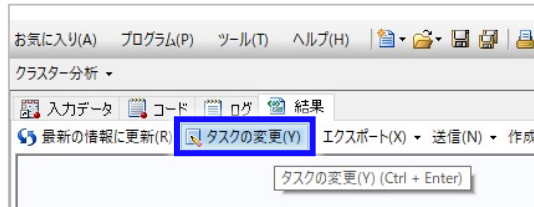


階層的クラスタリング（群平均法） – 実行結果

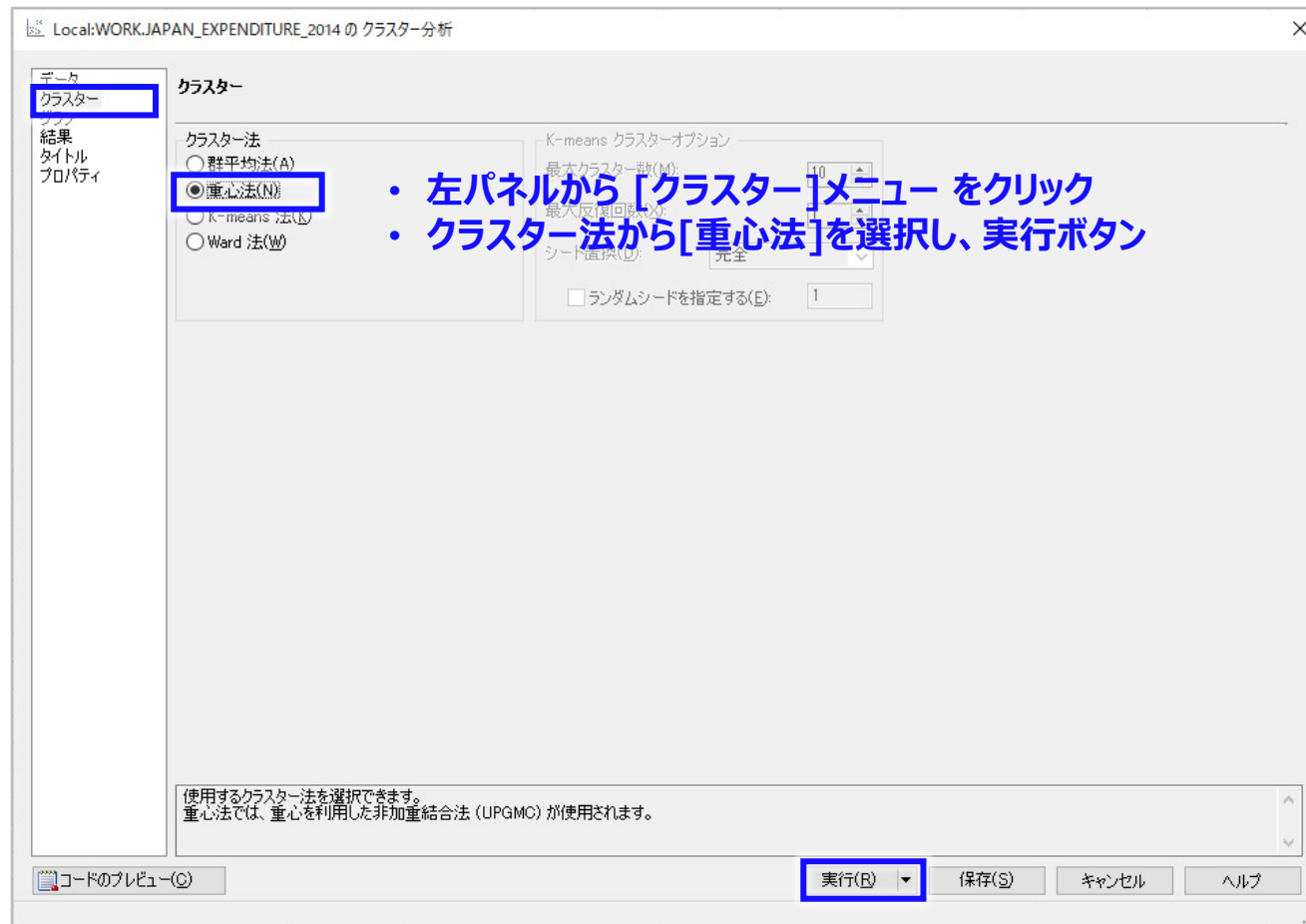
- 群平均法の結果、47都道府県が階層的にクラスタリングされた
- 任意の場所で区切ることで、最適なクラスタ数の検討が可能



階層的クラスタリング（重心法）：実行方法



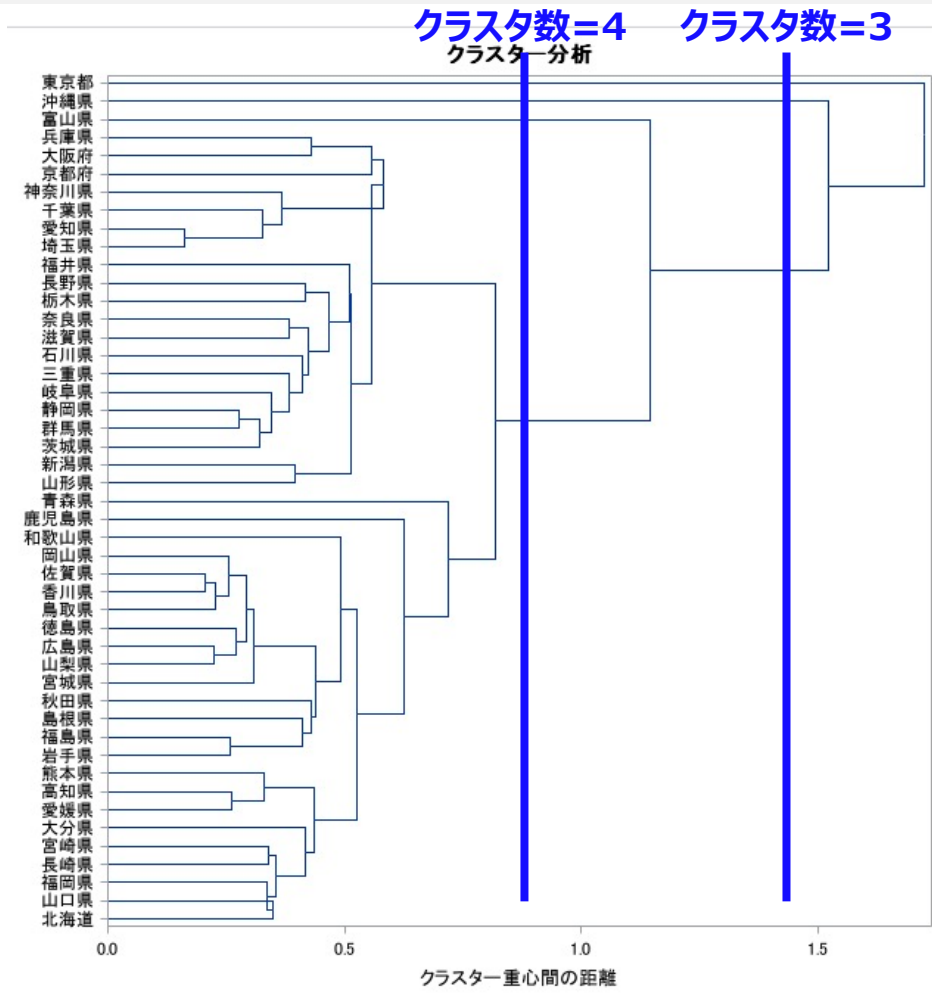
クラスタリング分析結果の画面上において
上部メニュー [タスクの変更] をクリック



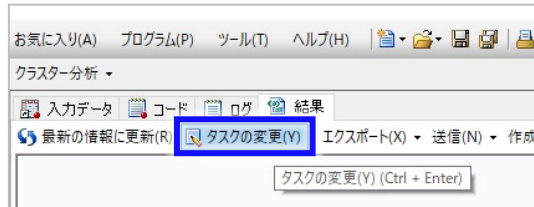
- ・ 左パネルから [クラスター] メニュー をクリック
- ・ クラスタリング法から [重心法] を選択し、実行ボタン

階層的クラスタリング（重心法） – 実行結果

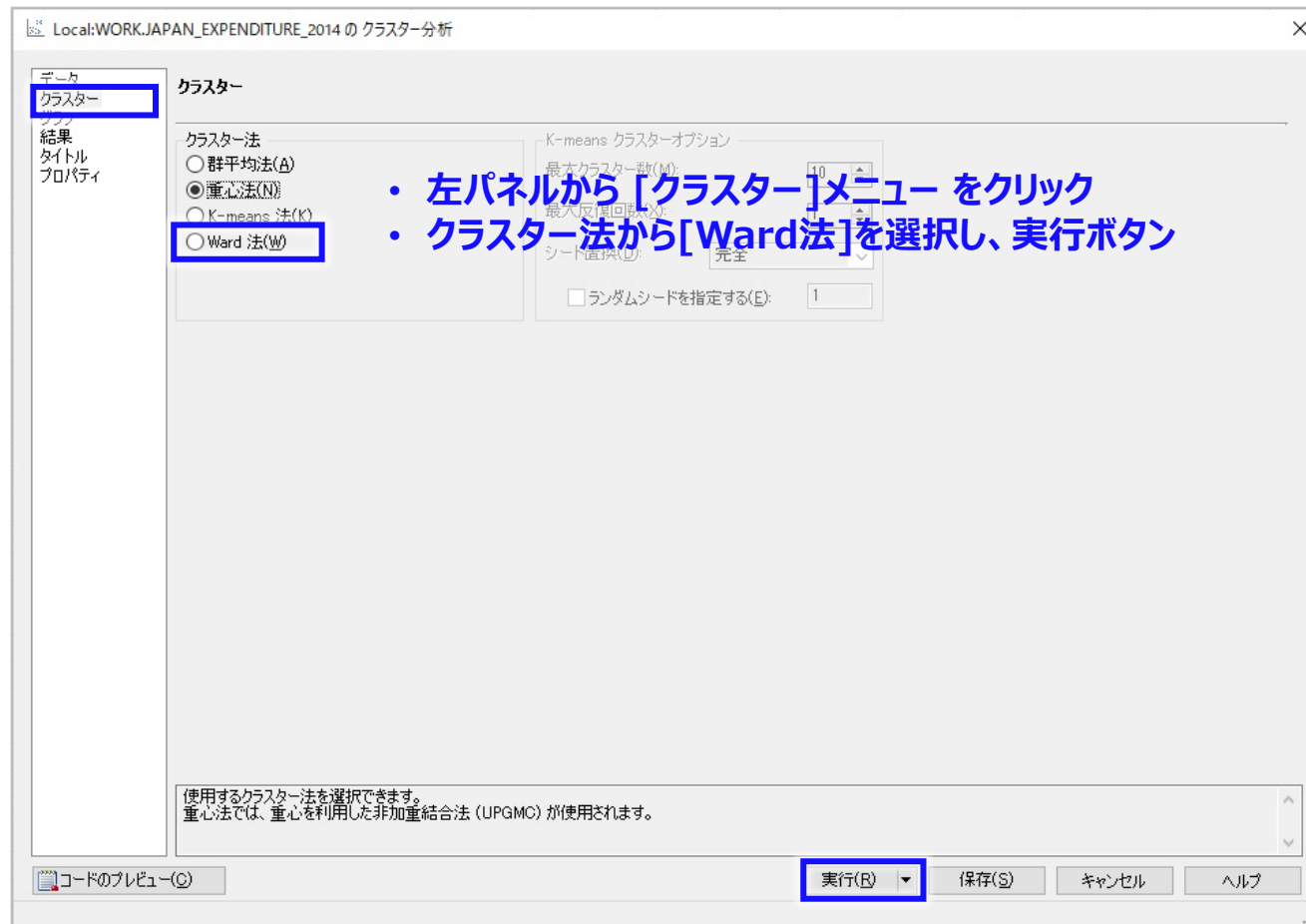
- 重心法の結果、47都道府県が階層的にクラスタリングされた。群平均法と類似の傾向を示すが、部分的に、分割結果が複雑化している
- 任意の場所で区切ることで、最適なクラスタ数の検討が可能



階層的クラスタリング (Ward法) : 実行方法

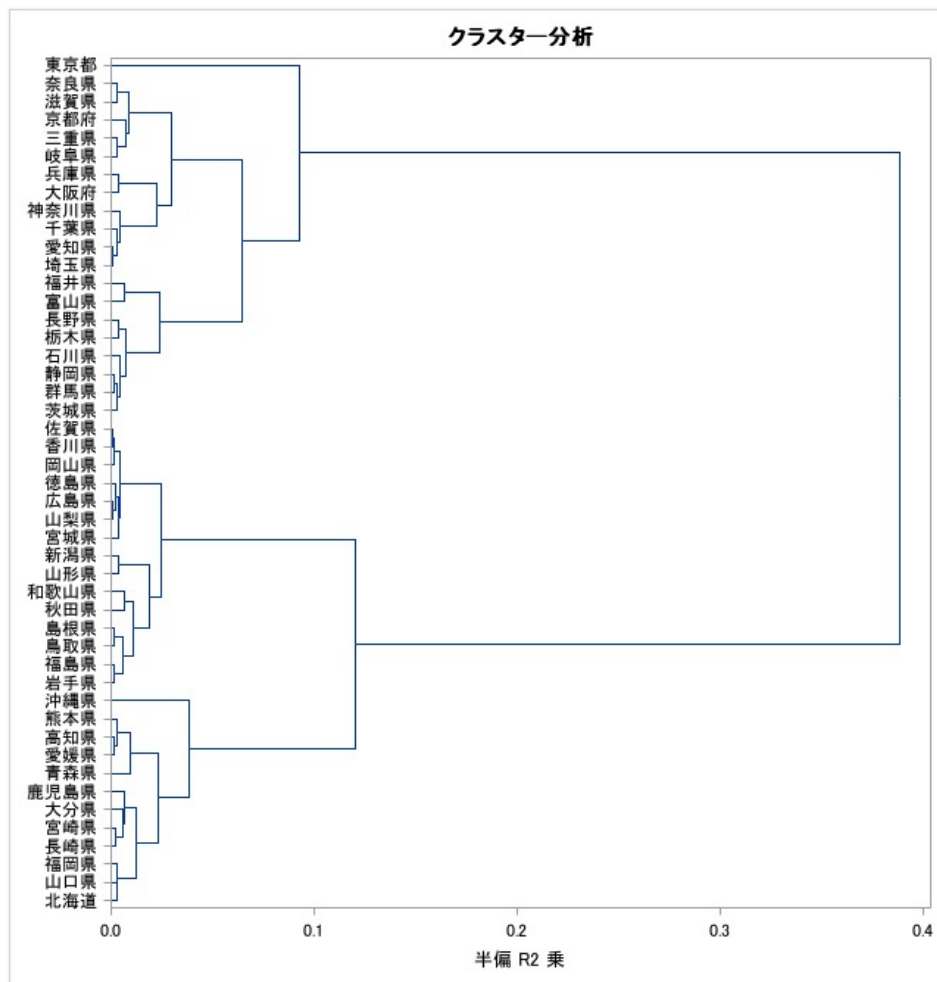


クラスタリング分析結果の画面上において
上部メニュー [タスクの変更] をクリック

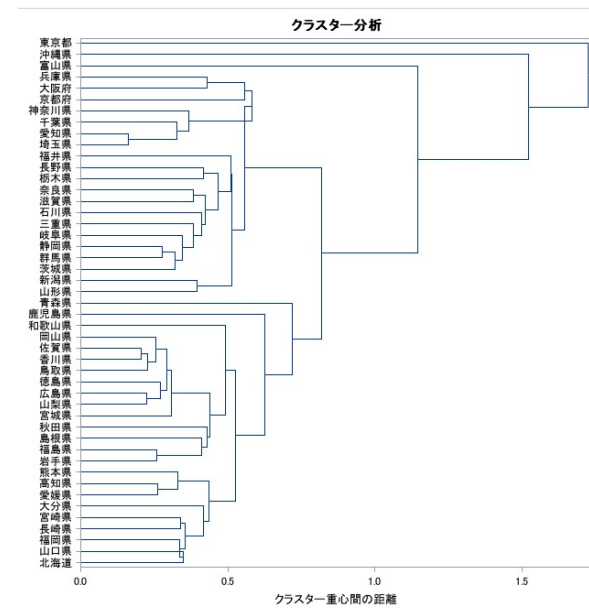


階層的クラスタリング (Ward法) – 実行結果

- Ward法では、全体的にバランスよく、分類が行われていることが確認できる

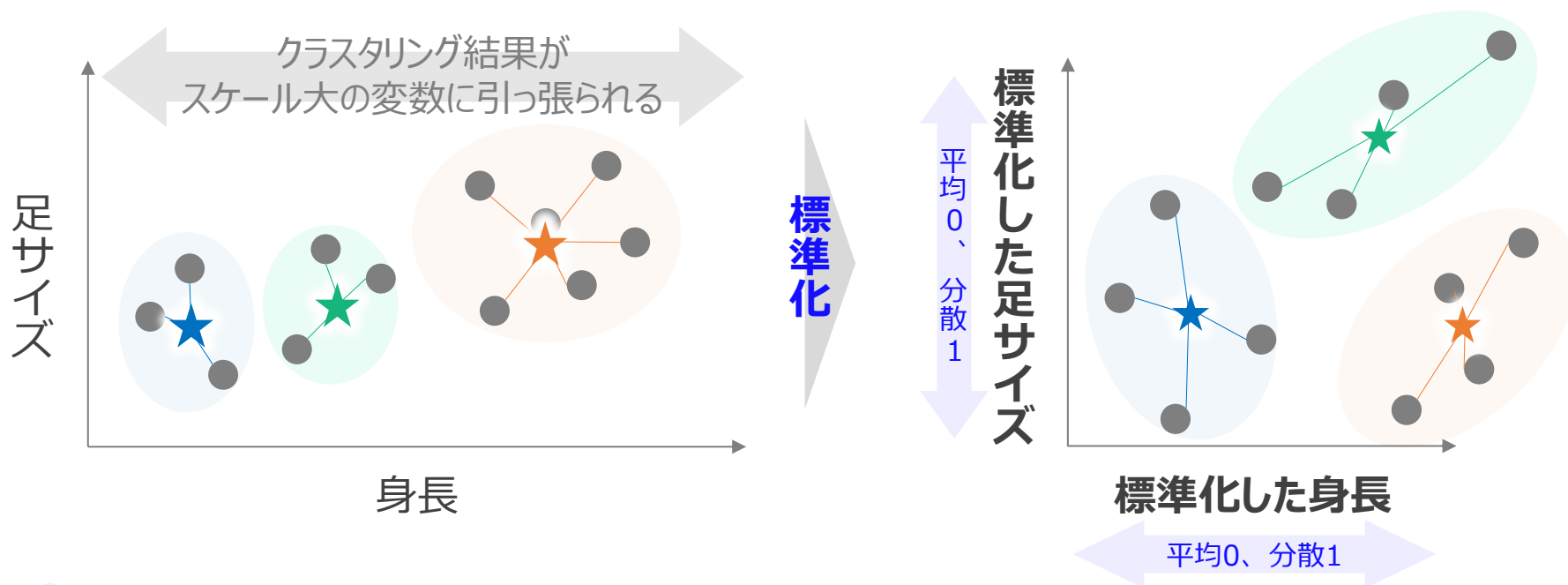


参考：重心法の結果



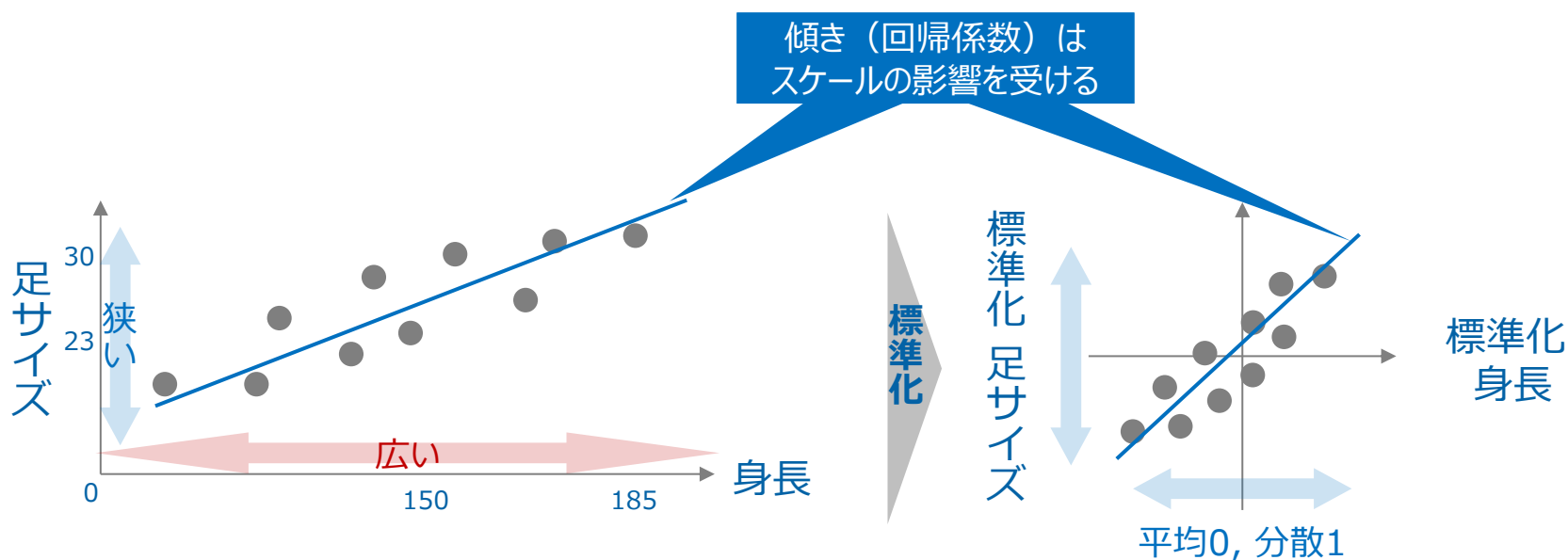
クラスタリングにおける変数スケールの影響と標準化

- k-means法などの「距離」に基づくクラスタリング手法は、データの「**スケール**」に大きく影響を受ける。このため、必要に応じて、「**標準化**」の処理を行なった上でクラスタリングを行う必要がある



(参考) 回帰分析における標準化の有効性

- 機械学習では、各変数間でスケール（値範囲）が大きく異なると、計算に時間がかかったり、回帰係数などのパラメータの直接比較が困難になるため、**スケールを揃える**ことが有効
- 特に、各変数を**平均0, 分散1**に変換する「標準化」を用いることが多い



データの標準化 – 実行方法 (1/2)

- [入力データ]タブ をクリック
- [データ]→ [データの標準化] をクリック

クラスター分析 ▾

入力データ | ログ | 出力データ | 結果

フィルタと並べ替え(L) | データ(D) ▾ | 記述統計(E) ▾ | グラフ(G) ▾ | 分析(Z) ▾ | エクスポート(X) ▾ | 送信(N) ▾

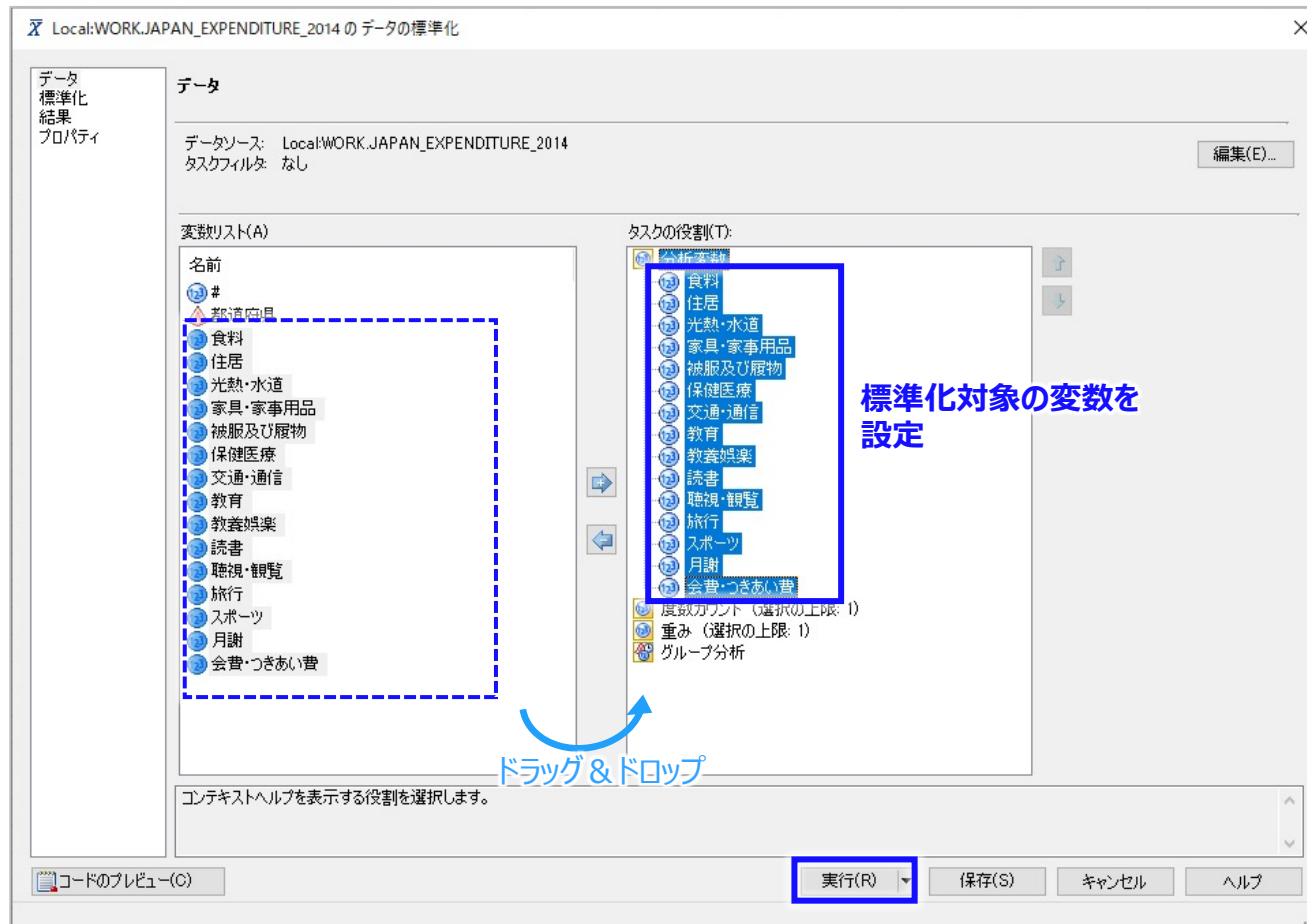
データの追加(B)...
データの並べ替え(S)...
出力形式の作成(C)...
データセットから出力形式を作成(M)...
転置(T)...
列の分割(P)...
列の積み上げ(K)...
ランダムサンプル(R)...
ランク(A)...
データの標準化(Z)...
データセットの属性(D)...
データの比較(O)...
データセットと出力形式の削除(N)...
LASR にアップロード(U)...
CAS にアップロード(U)...
データファイルをサーバーにアップロード(F)...

#	都道府県	食料	健康医療	交通
1	北海道	54281	10208	9234
2	青森県	55180	8972	9936
3	岩手県	57514	8288	10273
4	宮城県	59052	10640	10851
5	秋田県	59146	7467	9214
6	山形県	63042	9486	10407
7	福島県	57891	8949	9613
8	茨城県	62433	10645	10668
9	栃木県	63866	10252	12956
10	群馬県	61858	9682	11270
11	埼玉県	64632	11403	11616
12	千葉県	66536	11826	11859
13	東京都	68380	12404	12151
14	神奈川県	67197	11591	11443
15	新潟県	64400	9077	10628
16	富山県	67635	9387	10776
17	石川県	66478	9512	11135

データの標準化 – 実行方法 (2/2)

- 左側の[変数リスト]から、全変数を選択*し、右側の[分析変数]へドラッグ&ドロップにより移動

*全変数を選択してドラッグ&ドロップをすれば、自動的に数値型変数のみがセットされる



データの標準化 – 実行結果

- 標準化を行うと、元データの右側に、標準化された後のデータが出力される

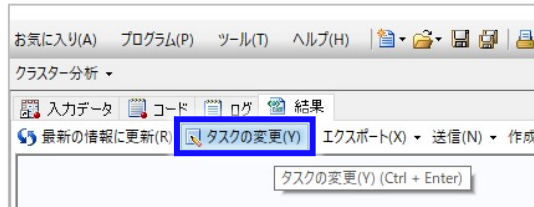
データの標準化 -

入力カテゴリー: コード ログ 出力カテゴリー

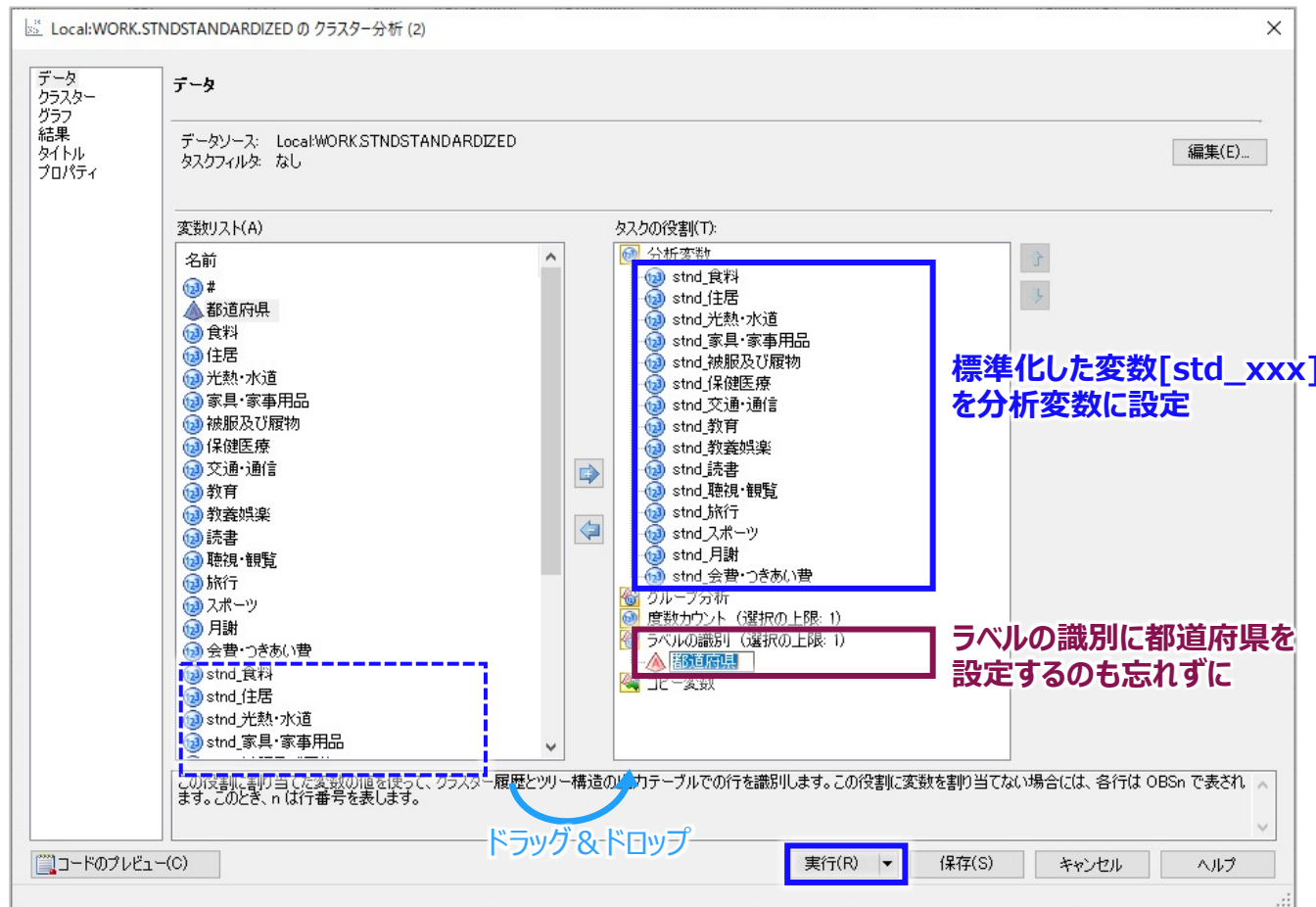
タスクの変更(Y) フィルタと並べ替え(L) クエリビルダ(Q) Where 式(W) データ(D) 記述統計(E) グラフ(G) 分析(Z) | エクスポート(O) | 送信(N) |

	観測値	旅行	スポーツ	月謝	会費・つぎの い	std. 食料	std. 住居	std. 光熱・ 水道	std. 家具・ 家事用品	std. 娯楽及 び贈物	std. 保健医 療	std. 交通・ 通信	std. 教育	std. 職業類 業	std. 読書	std. 娯楽・ 観覧	std. 旅行	std. スポーツ
1	2106	6506	2492	1430	101	-1.126440835	0.2824331478	0.4643504342	0.320673419	0.3941421031	-1.414683268	-0.774476972	-0.836208783	-0.370316738	0.916773616	-1.300082128	-0.403623942	0.6961091004
2	2084	3839	1153	777	107	-0.946537808	-0.523179513	1.8855653575	0.403847784	-0.489142484	-0.753005689	-1.36367215	-1.11588954	-2.246878089	-0.53543841	-1.337029273	-1.851479295	-2.106362369
3	2428	6826	1447	1219	145	-0.479470328	-0.413930827	1.3205099069	0.188505948	-0.977950459	-0.435362749	0.8840716269	-0.77341791	-1.215726299	0.0978436828	-0.75981028	-0.501341968	-1.491032562
4	2322	7734	2172	1587	116	-0.17169407	-0.010096274	0.8618005386	1.0168314604	0.7023629295	0.1094373802	0.3900358742	-0.208441495	-0.350604397	-0.569501326	-0.937328342	0.2630301167	0.0263623713
5	2403	6672	1617	979	119	-0.152883298	-0.904136106	1.8728234306	0.3423215414	-1.564662955	-1.433534481	-0.315044488	-1.266393247	-1.004990596	-0.323752331	-0.801295672	-0.313506211	-1.135229612
6	2227	5438	1659	1333	223	0.6267631901	-0.824191774	2.6745846769	3.3764768022	-0.121821871	-0.309059622	0.3042056997	-0.45427658	-0.022157924	0.4888490666	-1.05872831	-0.983417536	-1.047325554
7	1921	6701	1743	1432	102	-0.404027128	-0.980224603	0.7495735671	0.9404938445	-0.505579009	-1.05745278	1.1761674199	-1.214366081	-0.955301285	0.331791656	-1.610774028	-0.297762752	-0.871516837
8	2451	7254	2639	2041	113	0.5048933977	0.2312790574	0.8613104645	0.4311927807	0.7064360872	-0.063051291	1.3057095136	0.8215402727	0.5372838712	0.3340678504	-0.720683719	0.0024487258	1.003774004
9	2723	7996	2908	2400	99	0.7916576219	0.5687879115	0.6981157854	1.2606576812	0.425505891	2.0935274826	1.8621258176	-0.372084729	0.8063156822	0.1656294648	-0.263882655	0.4052641259	1.5667798482
10	2473	6800	2837	1965	93	0.3898275018	-0.196204623	-0.131089611	0.9085908484	0.0182459117	0.5043702217	1.0821412605	-0.131432914	0.27824705	0.627696928	-0.683736574	-0.24401784	1.4181797927
11	3115	8887	2699	2705	63	0.9449454067	0.9610602777	-0.404550965	-0.300285892	1.2431267965	0.3804962077	0.2557127894	1.873011587	1.2211077624	0.7028113432	0.3944482307	0.8889693328	1.1239515157
12	3512	10407	2889	2800	88	1.3259636083	0.3842272173	-0.261449325	0.0141815807	1.550415939	1.0595394463	0.0043692543	1.5812675406	1.590367111	1.581423815	1.0611763147	1.7141427372	1.5270136361
13	3882	14361	3061	2964	80	1.6949749337	4.3449419205	-1.106337093	-0.132797777	1.9634729706	1.334766157	-1.345691581	1.1721594553	2.0439037298	1.5677652151	1.68256001155	3.8606819447	1.8870002053
14	3607	10790	2874	2894	89	1.4582393599	1.6234928965	-0.791709514	-0.027975289	1.3824775265	0.6674332147	-0.134051829	1.1514263759	1.9923878152	1.5199651327	1.2207208041	1.922064972	1.4956192582
15	2535	7827	1955	1715	103	0.8985188191	-0.174611942	1.6214154114	0.1053315698	-0.414106172	-0.100753717	-0.010429052	-0.426730454	-0.216227311	0.2589534352	-0.579612388	0.0420788124	-0.427808629
16	4101	5869	2139	2433	137	1.5458895555	-0.48179954	1.6277863748	2.0696138341	-0.192570398	0.0387452592	2.845545632	-0.005553503	0.7188595207	1.1648788063	2.0503521402	-0.749437162	-0.04270526
17	3207	9516	1898	1966	140	1.3143569614	0.3305025695	-0.073260866	-0.084944032	-0.103241451	0.3771245336	0.6962469744	0.9366531226	0.8027063903	1.6588129911	0.5489545331	0.6875606327	-0.547108266
18	4031	9052	1863	2080	153	1.5046659475	-1.065074508	1.0627309234	0.2580078015	0.391283577	0.5203937528	1.490574505	0.6260649523	0.9237568134	-0.021018476	1.9327930514	0.4356652882	-0.620361814
19	3699	8404	1877	2212	119	-0.454055778	0.133698925	-0.178136726	-0.145437042	-0.162555869	-0.428764825	0.0828684428	0.4389137389	0.0905150496	1.3752270485	-0.4589974487	-0.501060939	
20	2910	7265	2537	1830	122	0.4934930385	1.2217147889	1.3611860582	1.2059676877	-0.201145972	1.1801862039	0.8690455547	0.0566457849	0.6813782385	1.2377170271	0.0501880768	0.0084203827	0.7902922341
21	2610	6754	2280	2056	130	0.4050367845	-0.93523985	0.6706624503	0.2671228004	0.2040501128	-0.256276225	0.5808201879	1.0403563074	0.3110084057	0.309029712	-0.453656626	-0.268990223	0.2524018923
22	2936	8355	2350	2259	104	0.4974891574	-0.345554003	-0.0811102052	0.2329415545	0.2347792691	0.7098484441	0.6192957834	0.0077749048	0.9509653969	0.5070556248	0.9038328845	0.6001572912	0.3909088993
23	3305	8547	2948	2655	81	0.8681013997	1.3091137374	-0.489823861	0.2306628048	0.9965764935	1.07938222	0.3028397022	1.3350622223	1.1968755649	1.2513741935	0.7135372695	0.7043898475	1.8504981893
24	3578	8754	2773	2385	117	0.6733899323	-0.909020165	0.3256694806	0.2602865512	1.2774266897	2.0303759189	0.2941863949	0.7878570186	1.056468068	0.1474199095	1.1720177495	0.8167655724	1.2842304458
25	2845	8377	1848	2052	148	0.6954024985	0.022927481	0.1149275927	0.3880821009	0.2637962147	0.0783328066	-0.1748393321	1.9428957816	0.5420037275	-0.480893745	0.6121006049	-0.651756192	
26	3109	8260	1855	2174	99	1.0860262009	-0.659804939	-0.315847551	-0.454101488	2.1249796992	-1.409970464	-0.543168066	1.5938554817	0.3762534786	1.1216311127	0.3849717967	0.5485038909	-0.637105482
27	3264	7492	2465	2250	65	0.4954801114	0.6182638736	-1.17608798	1.179724452	0.1726063249	0.0444006231	-1.817416042	0.9085531596	0.0997254251	0.5758444568	0.6468812267	0.1316536695	0.635992201
28	3023	8937	2546	2548	82	0.7424294299	0.7376788474	-0.905406707	-0.599941471	0.7543184006	0.1801298571	-0.680450812	0.7078865633	0.6505604232	0.784286552	0.2899420484	0.6229581629	0.8091288609
29	3065	9467	2170	2609	98	1.3003489389	0.3181638945	0.5937299996	1.2162220615	1.0087252297	1.5741765628	0.8114406946	2.4387248687	0.6841546947	0.8143448688	0.3104775059	1.2038375134	0.0221764542
30	2656	4955	2098	1787	80	-0.390213485	-1.464260544	-0.709377063	-0.746920824	-0.290474915	-2.270528341	-0.613744602	-0.700052132	-0.212895647	-0.262295082	-1.223912445	-0.12851556	
31	3787	7077	1775	1865	90	-0.376811537	-0.711087252	-0.041406049	-0.757175202	-0.433401223	-0.391062899	0.5785435255	-1.084043688	-0.123496016	-0.667457685	1.5230156261	-0.093640662	-0.804542165
32	3538	6335	1619	1512	165	-0.137474474	-1.148081895	0.4516085073	0.1224321928	-0.835641195	1.0171232169	0.3854825493	-1.491300785	-0.271755032	-0.102961474	1.1048411224	-0.496458662	-1.131043895
33	2796	7265	2630	1789	93	-0.308572458	-0.13059587	-0.594245596	0.1354454947	-0.365513261	-0.011567388	-0.2062200937	0.0714062871	-0.999782068	-0.14128531	-0.308077432	0.5663456716	
34	2660	7944	1918	2004	99	-0.370607984	0.341555896	-0.70790684	0.4243565315	-0.024631981	0.4336781728	-0.102178549	0.3254353004	0.009175158	-0.417076302	-0.369685943	0.3770344752	-0.505249095
35	2873	6234	1514	1798	84	-0.816063088	-0.51338616	-0.96176523	0.653370879	1.161620448	0.21311898	-0.797926595	-1.000008849	-0.201512465	0.149696104	-0.011970303	-0.55123673	-1.35003044
36	3782	7204	2039	1810	99	-0.803255753	-0.008942268	-0.273211104	-0.1453309	-0.000934508	-0.446678477	-0.118798184	-0.457238448	-0.203783573	-0.118894835	1.4810302341	-0.100658075	-0.252001119
37	3814	8103	2606	2126	84	-0.511000000	-0.245000000	-0.140000000	-0.500000000	-0.440000000	-0.310000000	-0.400000000	-0.200000000	-0.200000000	-0.200000000	-0.200000000	-0.200000000	-0.200000000

階層的クラスタリング (Ward法) : 標準化 – 実行方法



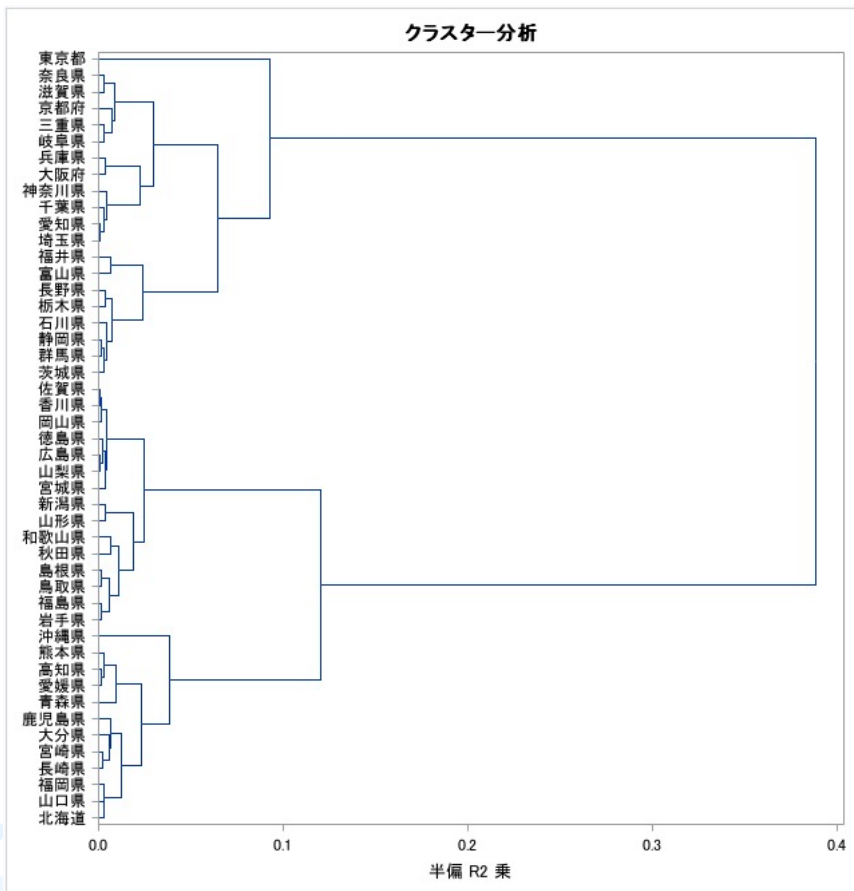
クラスタリング分析結果の画面上において
上部メニュー [タスクの変更] をクリック



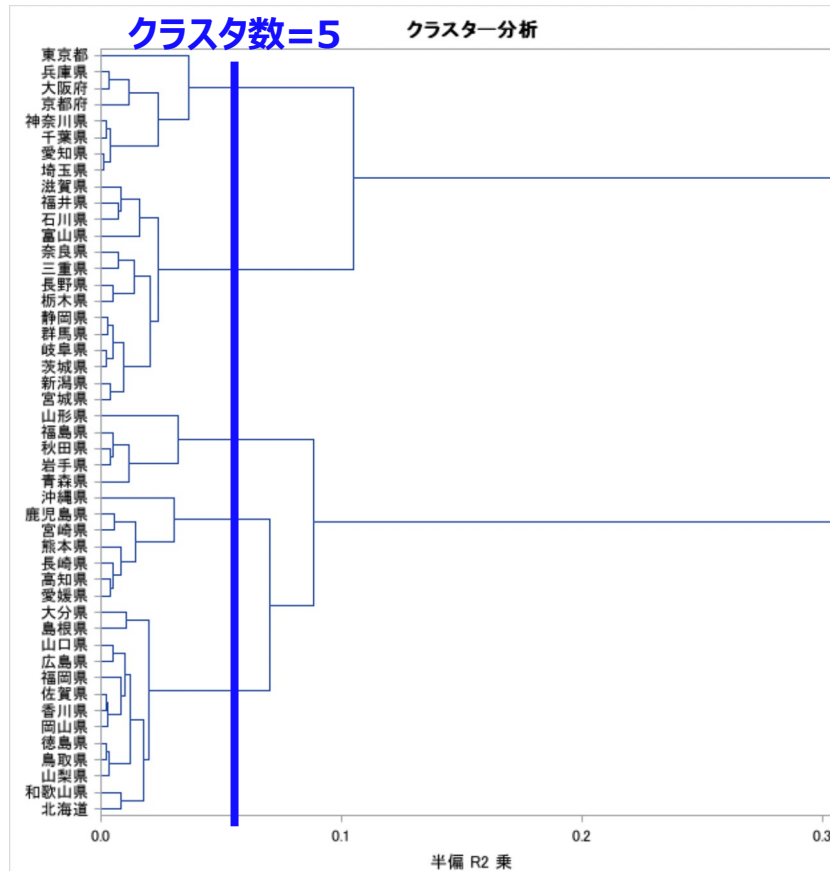
階層的クラスタリング (Ward法) : 標準化 - 実行結果

- 標準化を行うことで、全ての変数がフェアに分類に寄与するようになり、分類結果が変わることが確認できる

標準化前

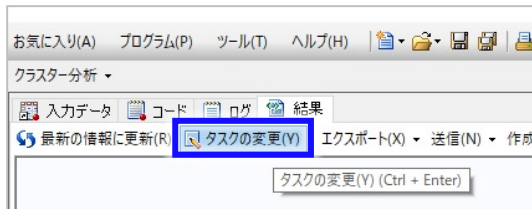


標準化後



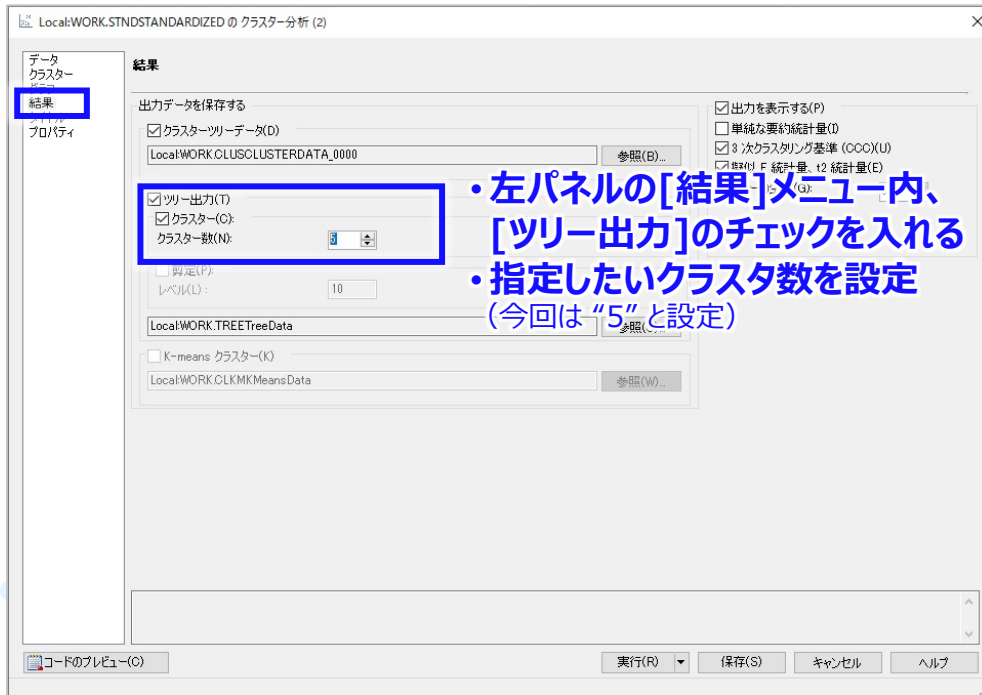
指定したクラスタ数の分類結果を出力する方法

- 階層的クラスタリングでは、分析の試行錯誤を重ね、結果の妥当な解釈を行いながら、最適なクラスタ数を決めていく
- 最適なクラスタ数の決定後は、明示的にクラスタ数を指定した分類結果を出力することが可能



クラスタリング分析結果の画面上において
上部メニュー [タスクの変更] をクリック

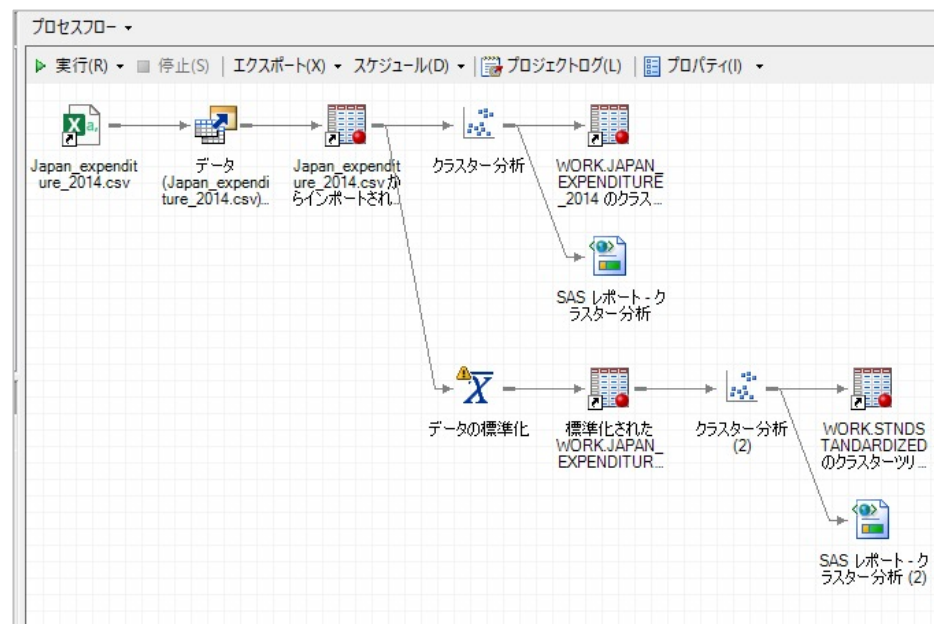
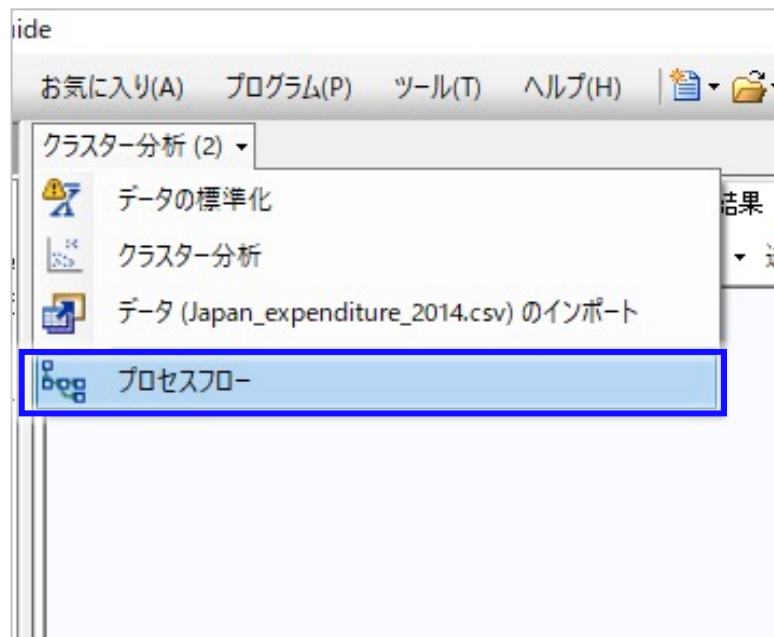
指定したクラスタ数での
分類結果が出力される



	CLUSTER	CLUSNAME
1	埼玉県	CL5
2	愛知県	CL5
3	香川県	CL11
4	佐賀県	CL11
5	鳥取県	CL11
6	徳島県	CL11
7	千葉県	CL5
8	神奈川県	CL5
9	茨城県	CL9
10	岐阜県	CL9
11	岡山県	CL11
12	群馬県	CL9
13	静岡県	CL9
14	大阪府	CL5
15	兵庫県	CL5
16	山梨県	CL11
17	岩手県	CL6
18	秋田県	CL6
19	愛媛県	CL7
20	高知県	CL7
21	宮城県	CL9
22	新潟県	CL9
23	広島県	CL11
24	山口県	CL11
25	福島県	CL6
26	栃木県	CL9
27	長野県	CL9
28	長崎県	CL7
29	宮崎県	CL7
30	鹿児島県	CL7
31	石川県	CL9
32	福井県	CL9
33	三重県	CL9
34	奈良県	CL9
35	滋賀県	CL9

(参考) プロセスフローでの確認

- SAS Enterprise Guideでは、「プロセスフロー」によりデータ加工やグラフ観察、モデル構築などの一連の分析プロセスを俯瞰的／反復的に確認可能

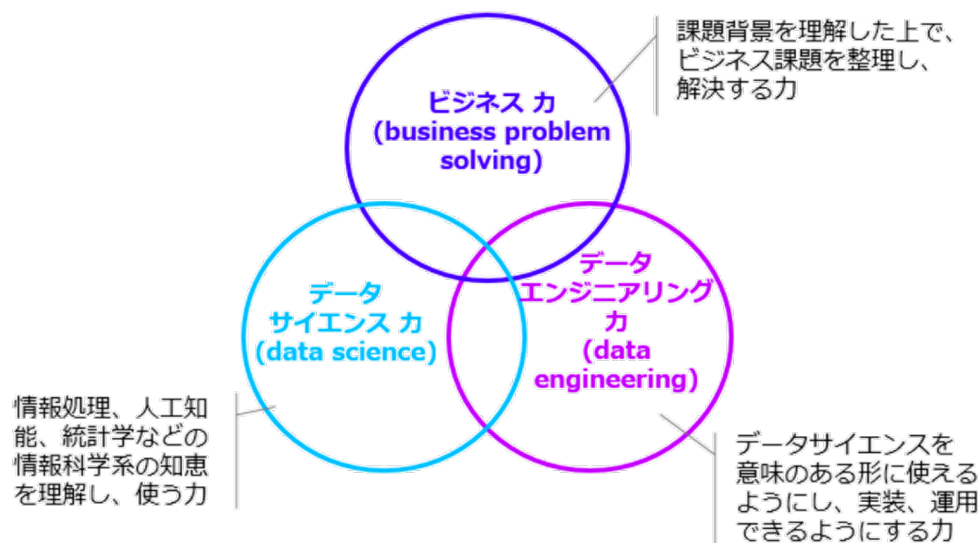
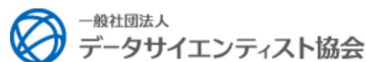


Agenda

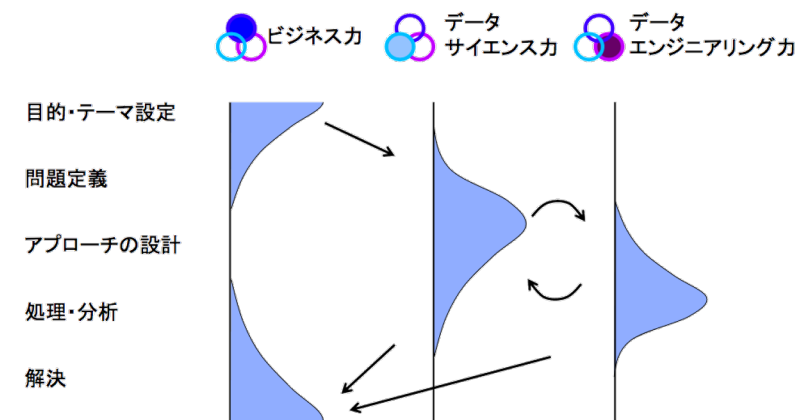
- クラスター分析による分類（2）：階層的クラスタリング
 - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
 - 各クラスターの解釈、予測への適用方法
 - 樹形図（デンドログラム）とクラスター数の検討
 - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する
- 今後のデータサイエンス学習に向けて
 - データサイエンティストに求められるスキル
 - 学習リソースの紹介
 - SAS内サンプルデータの紹介と使い方

データサイエンティストに求められる知識・スキルセット

- データサイエンティスト協会が定義するデータサイエンティスト：
「データサイエンティストとは、**データサイエンス力**、**データエンジニアリング力**をベースにデータから価値を創出し、**ビジネス課題**に答えを出すプロフェッショナル」
- これら3スキルはどれも不可欠で、分析フェーズによって中心となるスキルが変化する、としている



課題解決の各フェーズで要求されるスキルセットのイメージ

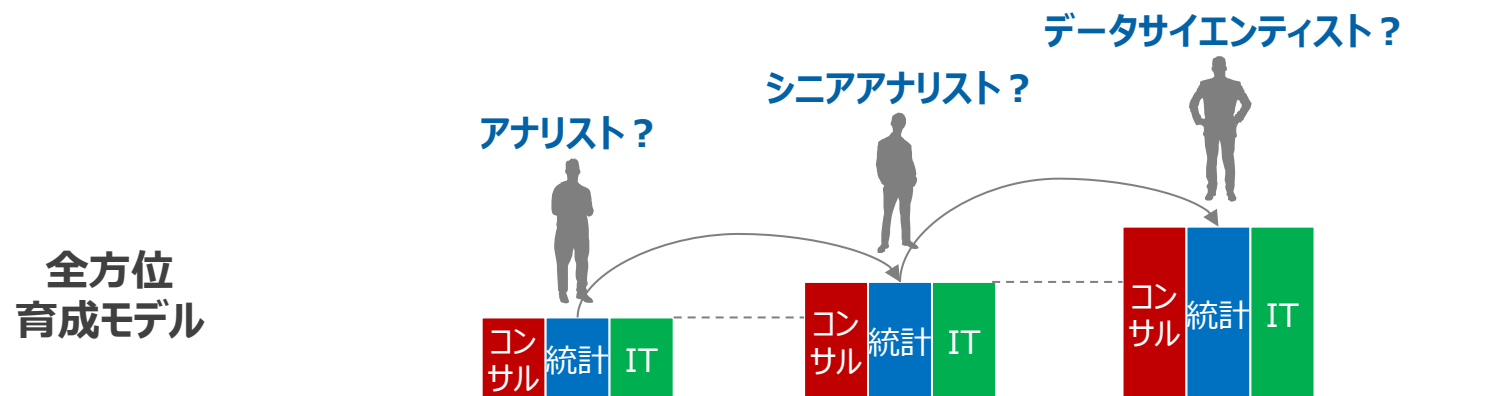


Source: The Japan Data Scientist Society discussions

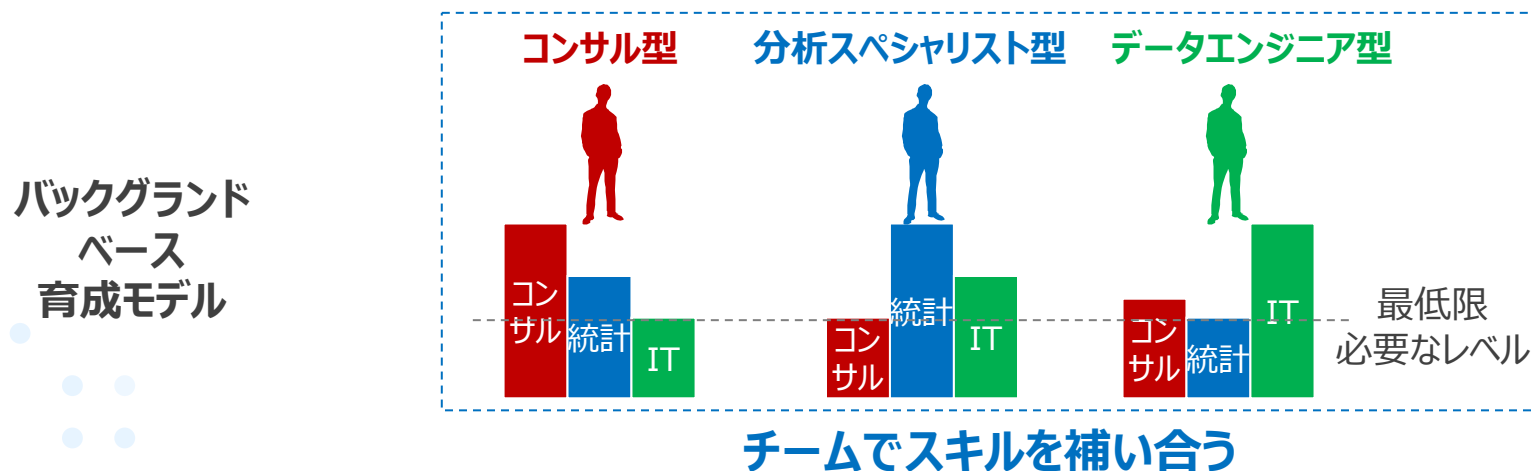
出展：データサイエンティスト協会資料
<http://www.datascientist.or.jp/news/2014/pdf/1210.pdf>

データサイエンティストの育成モデルの例

- 理想的には、すべてのスキルを持つ人材を育成できればよいが、元々のバックグラウンドや経験値を生かした育成を行うのが現実的
- チームでスキルを補い合い、プロジェクトフェーズに応じて、役割分担や協業体制が必要



現実的には・・・



参考：データサイエンティスト協会が定義するスキルレベル

- バックグラウンドベース育成モデルを実現する上でも、各メンバーは、各スキルについて最低限のレベルを保有する必要がある。データサイエンティスト協会が定義する「**独り立ちレベル**」がまず目標

レベル定義	コンサル力 (ビジネス力)	統計力 (データサイエンス力)	IT力 (データエンジニアリング力)
業界代表	- 組織や市場全体にインパクトを出せる - 対象とする事業全体、産業領域における課題の切り分け、テーマ、論点の明確化ができる	- 新しいアルゴリズムや分析手法の開発ができる - 複数のパラメータやアルゴリズムの選択など、適切な分析アプローチの設定ができる	複数のデータソースを統合したデータシステム、もしくはデータプロダクトの構築、全体最適化ができる
棟梁	- 分析を通じオペレーション上の革新が実現できる - 仮説や可視化された問題がない中で、適切に問題を定義し、解き、価値を見出すことができる - 特定の課題領域において、課題と取組のテーマを構造的に整理し、見極めるべき論点をクリアにできる - 組織全体を見渡して、必要なデータの当たりをつけることができる	- 多変量解析の概念を理解し、活用することができる - 機械学習、自然言語、画像処理のアルゴリズムを理解し、適切に活用、問題解決することができる - モデルを構築できる	- 分析に必要なデータフォーマット、取得蓄積仕様等を設計できる - 問題設定に応じた新規データマート設計ができる - 構造化データ／非構造化データを問わず、分析システムを設計できる - 構築したモデルを実装できる - データ分析を作ったシステムを自身で構築できる
全てのスキルで「 独り立ちレベル 」以上の習得が理想			
独り立ち	- 仮説や既知の問題が与えられた中で、最適解・最大解を見出すことができる - 扱っている課題領域で新規の課題を切り分け、構造化できる - 当該プロジェクト・サービスを超えて、必要なデータの当たりをつけることができる	- SPSS/SAS/R等が使える。指示されなくてもサンプル抽出ができるとともに内容を確認できる - データクレンジング、分布、単回帰やP値の概念を理解し、活用することができる	大規模のファイルや、データベースにアクセスし大量の構造化データを処理することができる
見習い	- ビジネスにおける論理とデータの重要性を認識している - 仮説や既知の問題が与えられた中で、必要なデータに当たりをつけて、データを用いて改善することができる - 扱っている課題領域における基本的な課題の枠組みが理解できる	基本統計量（平均、中央値など）の知識を有し、指示されればデータの抽出、グラフ作成を正しく行うことができる	- 一般的なアクセス解析システムを使うことができる - 抽出されたデータサブセットに対し、ExcelやAccess等の統合環境を用い、目的に応じた処理をすることができる
未経験	- ビジネスは勘と経験だけで回すものと思っている - 課題を解決する際に、そもそも定量化する意識がない	- 基本統計量の意味を正しく理解していない - 指数を指数で割り算したりする 「平均年収」をそのまま鵜呑みする - グラフ・チャートの使い方が不適切	- レポートされてくる数値サマリーに目は通すが、特に記憶には残らない - アクセス解析システムを使っていない - ExcelやAccessは数字しか入れない

今後のデータサイエンス学習に向けて (スキルアップ編)

- おすすめのオンライン学習サイト

- Gacco: <https://gacco.org/>

- 国（総務省）がやっている無料の講座集。データサイエンス学部がある滋賀大の講座が受講可能

- Coursera: <https://www.coursera.org/learn/machine-learning>

- 世界中でもっとも有名かつ人気と言っても過言ではない、スタンフォード大の機械学習講座。無料です
 - 日本語字幕もつけられます

- 東大松尾研の無料講座: <https://gci.t.u-tokyo.ac.jp/>

- 日本のデータサイエンスでもっとも有名とも言える東大松尾先生の無料講座
 - Pythonベースなので、Pythonの学習にも良いと思います
 - 年数回（不定期？）受講生を募集しているので、こまめにチェックしてください

- 今後の分析技術

- テキスト分析、画像分析

- スペースモデリング、ベイズ統計、xAI (eXplanable AI)

今後のデータサイエンス学習に向けて (腕試し／リアルデータ編)

• データ分析コンペ

– **Kaggle:** <https://www.kaggle.com/>

- Kaggleは本場アメリカのコンペサイト
- ここで上位に入り込めれば、AI系の有力企業からも注目を浴びると思います

– **SIGNATE:** <https://signate.jp/>

- "日本版Kaggle" とも言え、規模は小さいですがまずはこちらから挑戦いただくのが良いかもしれません

• プロジェクトベース

– **AI Quest** (経産省主催、BCG運営の課題解決型データサイエンスコース)
<https://aiquest.meti.go.jp/2021/>

• オープンデータを活用した独自分析

– データカタログサイト: <https://www.data.go.jp/>

– 観光統計データ: <https://statistics.jnto.go.jp/>

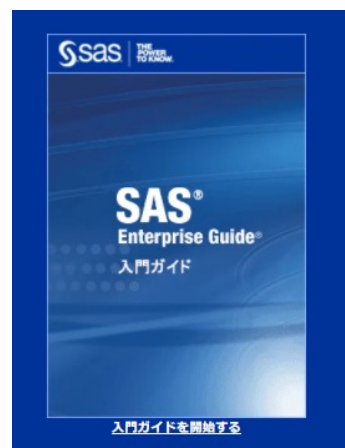
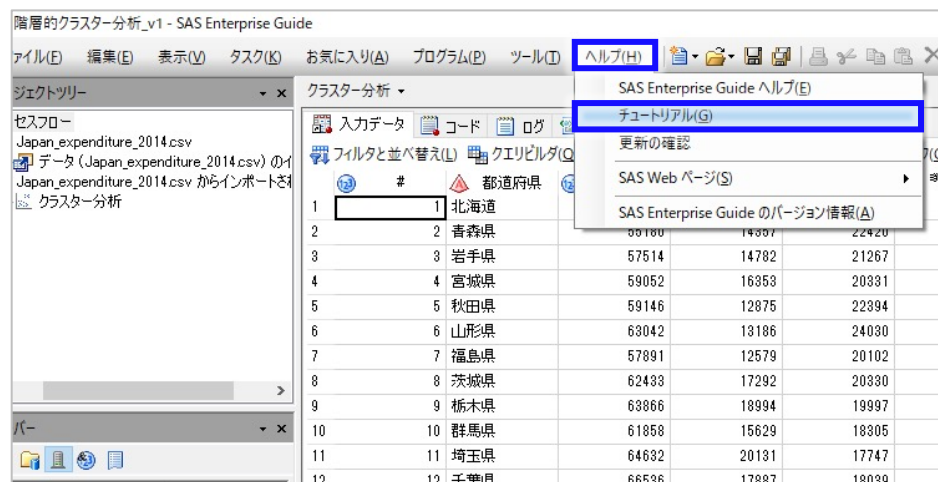
– 政府統計e-stat: <https://www.e-stat.go.jp/>

– 医療: NDBオープンデータ: <https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000177182.html>

SAS Enterprise Guide チュートリアルを活用

- SAS Enterprise Guideの使い方などに困った場合は、オフィシャルのチュートリアルを閲覧すると有益な情報が得られることがある

- 上部メニュー[ヘルプ]から、[チュートリアル]をクリック

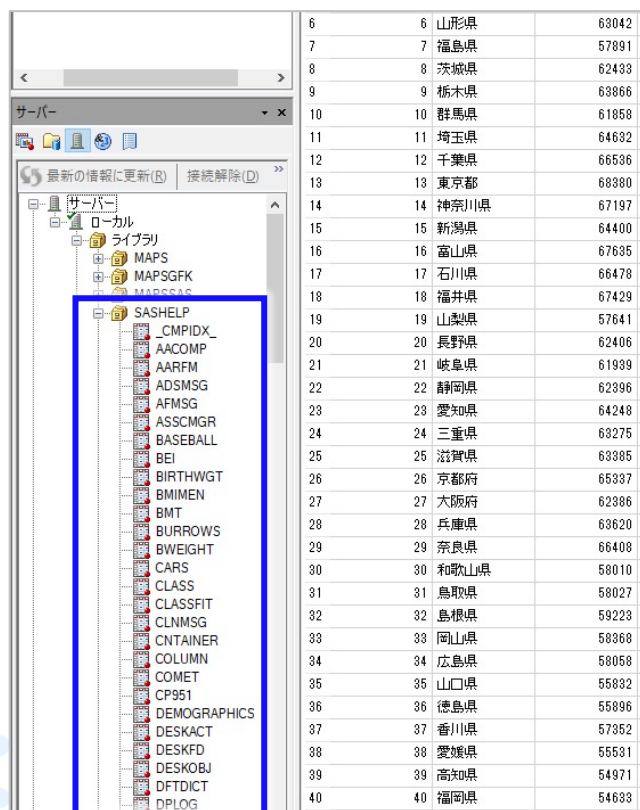


- 概要
- プロジェクトを開始してメインウィンドウを開く
- SASデータのプロジェクトへの追加
- SASライブラリのデータの探索
- テキストファイルからのデータのインポート
- SASタスクについて
- リストの作成と変更
- 棒グラフの作成
- クエリビルダについて
- クエリによるテーブルの結合
- クエリへの計算列の追加
- クエリからの要約テーブルの作成
- 円グラフの作成
- 線形モデル分析の実行
- 複数レポートを単一ドキュメントに結合する
- プロセスフローの操作
- その他の機能を習得するには

SASHELPデータ (サンプルデータ) の活用

- SAS Enterprise Guideでは、デフォルトで様々なサンプルデータ (**SASHELP Data Sets**) が格納されており、分析のトレーニングなどに有効活用できる

- ウィンドウ左下の[サーバー]のパネルから
[サーバー]→[ローカル]→[ライブラリ]→[**SASHELP**]を選択



▼データセットの例

BASEBALL	1986年のメジャーリーガーの成績データ
BMIMEN	年齢とBMIに関するデータ
BMT	骨髄移植患者の生存期間データ
BIRTHWGT	2003年の乳児死亡率に関するデータ
FAILURE	機械の不具合に関するデータ
DEMOGRAPHICS	各国の人口などに関するデータ
JUNKMAIL	迷惑メールデータ
ORSALES	売上に関するデータ

データセットの一覧と詳細は、下記リンクを参照 (英文)

<https://support.sas.com/documentation/tools/sashelpug.pdf>

まとめ

- 階層的クラスタリングによるデータ分類
 - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみについて学習した
 - 各手法を都道府県データに適用し、類似の都道府県をグルーピングすることができた
 - デンドログラムを観察することで、最適なクラスタ数を検討することができた
 - データを標準化することで、さらに精緻なクラスタリングを行うことができた
- 今後のデータサイエンス学習に向けて
 - 無料の有用な学習リソースはたくさん存在
 - 実践的なスキルを鍛えるには、座学だけでなく、実際のデータに触ってみることが一番
 - データサイエンスの知識だけでなく、「ビジネス力」「IT力」も極めて重要

アンケートのお願い・ご質問

機械学習によるビッグデータ分析の手法

今後の参考にさせていただくため、ぜひともアンケートにご協力をお願いします。

- ・無記名
- ・所要時間目安：1 ～ 3 分

アンケートURL

https://sas.qualtrics.com/jfe/form/SV_e9byCaaw8q5xob4

- ・お客様講演会のアーカイブは、2022年3月31日迄視聴できます。

本日の内容に関するご質問は、以下宛にご連絡ください。

que@datascience.co.jp



End of File