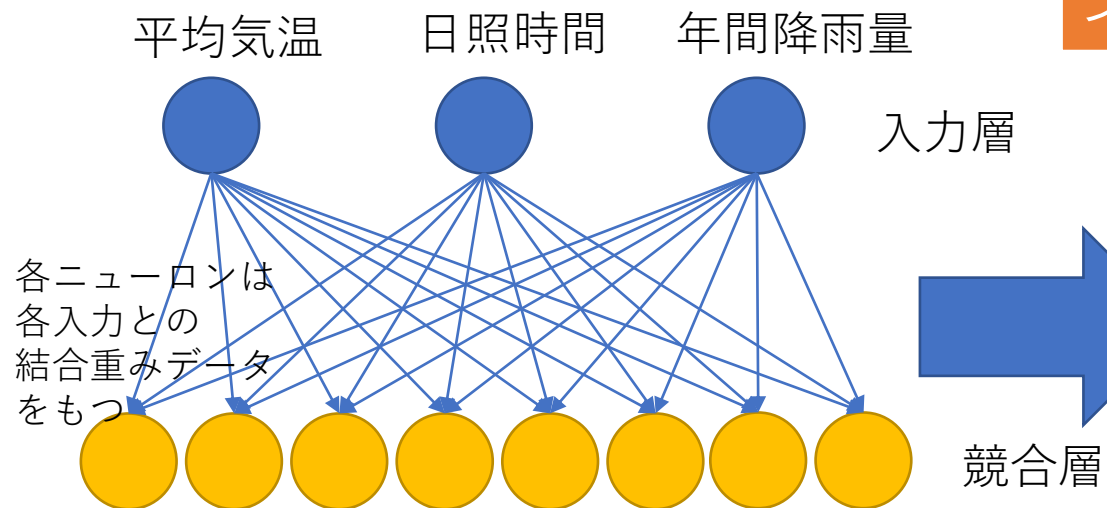


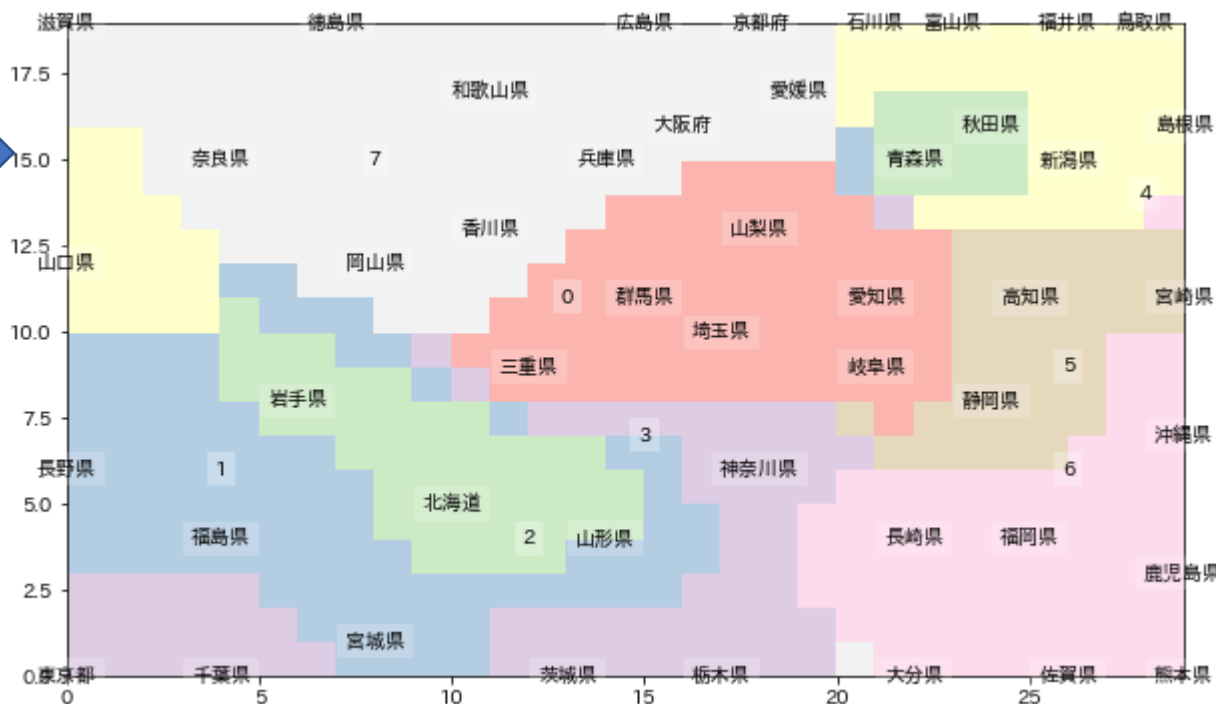
自己組織化マップ (SOM) による都道府県データの分析

実験概要



課題の概要

- 分析対象のデータは公開されている統計データから選び自分で作る
- SOM + K-means法でクラスタリングし、結果を分析



東北・関東・関西・日本海側という概念に似た結果

実験の手順（概要）

1. 添付されたデータでSOMを使ったデータ分析手順を理解する。
クラスタリングの結果をレポート（例なので分析は不用）
2. Webで公開されているデータを加工して、
独自の実験用データを作成
入手先やどの項目を使ったかを正確にレポートに記述
3. 独自データでデータ分析を実行
実行前の予想、結果の図、結果の分析をレポート記述
乱数が使われているのでクラスタリング結果は1通りではない。
同じデータに対して複数回実行してみよう

サンプルによるデータ分析手順の理解

1. SOM.zipを解凍する。SOMというフォルダができる。
2. Chromeブラウザで[Google Colaboratory](https://colab.research.google.com/)にログイン
3. SOMex.ipynb をアップロード
4. 説明通りに実行



レポートについて

- レポートは、共通教材の、[レポートについての注意事項](#) に従って作成してください。
- 実験用データは、csv形式保存できるなら何を使ってもOK。Excelでなくても構いません。
- Google Colabolatory による実行なので自宅のPCでも実行できます。
頑張ればスマホだけでもできなくはないと思いますが、PCを使う方が楽でしょう。
- PCを持っていないという人は、大学から貸し出しを受けるか、大学の自由利用PC教室を使ってください。
- レポートは原則Wordで作成してください。フリーのOfficeで作成しても構いませんが、その場合はPDFで出力し、正しく表示されることを確認してから提出してください。

都道府県データの入手

<https://uub.jp/pdr/>

の例で説明する

他のサイトのデータでも構いません

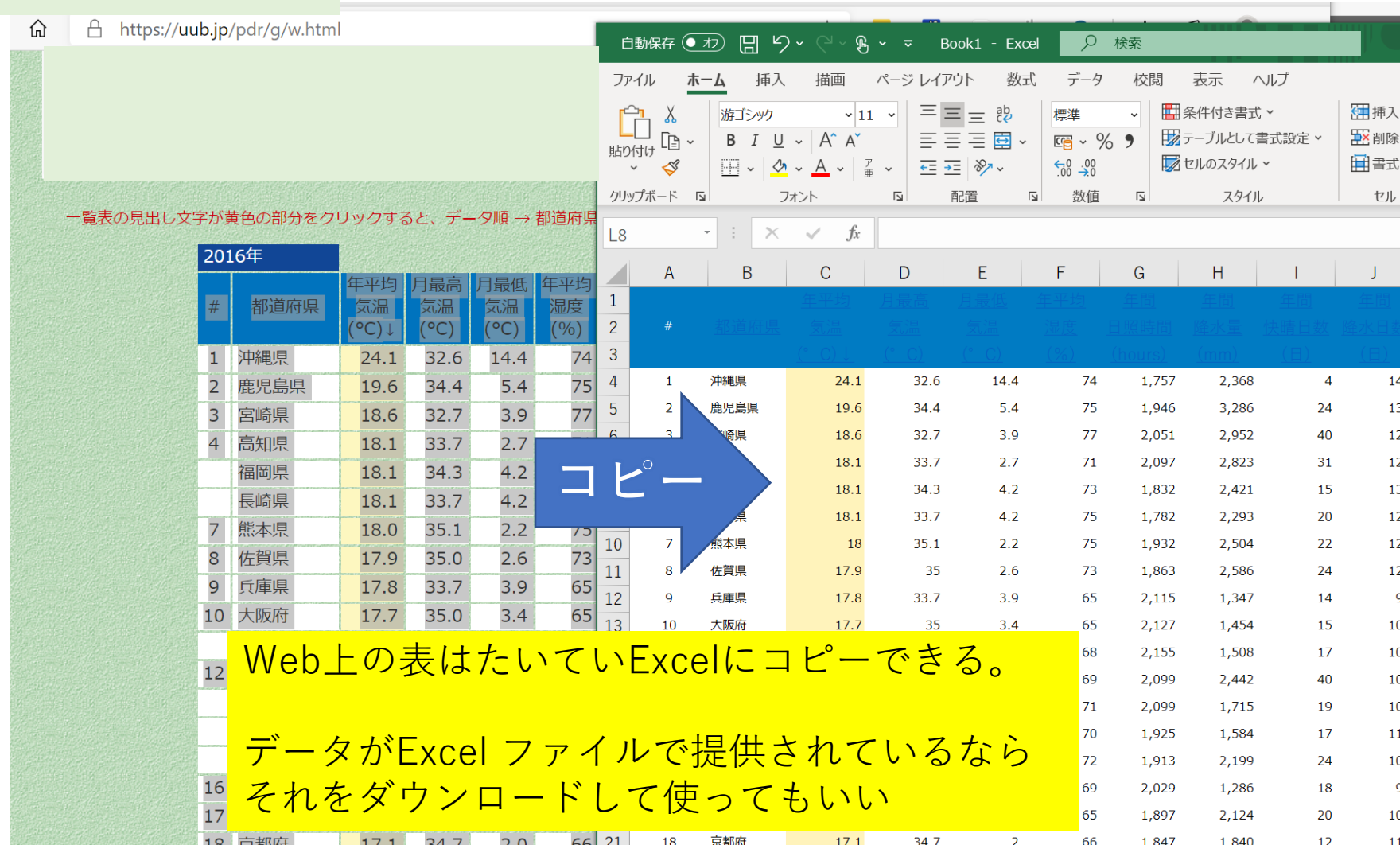
一覧表の見出し文字が黄色の部分をクリックすると、データ順 → 都道府県

2016年					
#	都道府県	年平均 気温 (°C)↓	月最高 気温 (°C)	月最低 気温 (°C)	年平均 湿度 (%)
1	沖縄県	24.1	32.6	14.4	74
2	鹿児島県	19.6	34.4	5.4	75
3	宮崎県	18.6	32.7	3.9	77
4	高知県	18.1	33.7	2.7	
	福岡県	18.1	34.3	4.2	
	長崎県	18.1	33.7	4.2	
7	熊本県	18.0	35.1	2.2	75
8	佐賀県	17.9	35.0	2.6	73
9	兵庫県	17.8	33.7	3.9	65
10	大阪府	17.7	35.0	3.4	65
12					
16					
17					
18	京都府	17.1	34.7	2.0	66

コピー

Web上の表はたいていExcelにコピーできる。

データがExcel ファイルで提供されているならそれをダウンロードして使ってもいい



データの加工

プログラムで扱いやすい形に編集する

	A	B	C	D	E	F	G	H	I	J	K	L
1		都道府県	年平均	月最高	月最低	年平均	年間	年間	年間	年間	年間	
2			気温	気温	気温	湿度	日照時間	降水量	快晴日数	降水日数	雪日数	
3			(°C) ↓	(°C)	(°C)	(%)	(hours)	(mm)	(日)	(日)	(日)	
4	1	沖縄県	24.1	32.6	14.4	74	1,757	2,368	4	145	0	
5	2	鹿児島県	19.6	34.4	5.4	75						
6	3	宮崎県	18.6	32.7	3.9	77						
7	4	高知県	18.1	33.7	2.7	71						
8		都道府県平均	16.2	32.7	1.4	71						
9		福岡県	18.1	34.3	4.2	73						
10		長崎県	18.1	33.7	4.2	75						
11	7	熊本県	18	35.1	2.2							
12	8	佐賀県	17.9	35	2.6							
13	9	兵庫県	17.8	33.7	3.9	65						

	A	B	C	D	E	F	G	H	I	J	
1	都道府県	平均気温	最高気温	最低気温	平均湿度	日照時間	降水量	快晴日数	降水日数	雪日数	
2	沖縄県	24.1	32.6	14.4	74	1,757	2,368	4	145	0	
3	鹿児島県	19.6	34.4	5.4	75	1,946	3,286	24	132	5	
4	宮崎県	18.6	32.7	3.9	77	2,051	2,952	40	122	3	
5	高知県	18.1	33.7	2.7	71	2,097	2,823	31	125	6	
6	福岡県	18.1	34.3	4.2	73	1,832	2,421	15	134	17	
7	長崎県	18.1	33.7	4.2	75	1,782	2,293	20	127	11	
8	熊本県	18	35.1	2.2	75	1,932	2,504	22	127	12	
9	佐賀県	17.9	35	2.6	73	1,863	2,586	24	121	11	
10	兵庫県	17.8	33.7	3.9	65	2,115	1,347	14	96	7	
11		17.7	35	3.4	65	2,127	1,454	15	105	9	
12		17.7	33.9	3.6	68	2,155	1,508	17	100	7	
13		17.6	32.3	2.3	69	2,099	2,442	40	107	1	
14		17.6	33.9	3.5	71	2,099	1,715	19	105	8	
15		17.6	34	3.2	70	1,925	1,584	17	111	11	
16		17.6	33.9	2.8	72	1,913	2,199	24	102	8	

- 不要な行、列の削除
- 見出し行を 1 行に修正

数値書式の編集 桁区切りカンマの削除

- 統計データは千単位でカンマ区切りされていることが多いが、プログラムでの処理には邪魔なので除去しておく必要がある



The screenshot shows the Excel ribbon with the '数式' (Formulas) tab active. The '条件付き書式' (Conditional Formatting) dropdown is open, and the '標準' (Standard) option is selected. The spreadsheet below shows a table of weather data with columns for average temperature, maximum temperature, minimum temperature, average humidity, sunshine hours, and precipitation. The values are formatted with commas removed.

県	平均気温	最高気温	最低気温	平均湿度	日照時間	降水量	快晴
	24.1	32.6	14.4	74	1757	2368	
	19.6	34.4	5.4	75	1946	3286	
	18.6	32.7	3.9	77	2051	2952	
	18.1	33.7	2.7	71	2097	2823	
	18.1	34.3	4.2	73	1832	2421	
	18.1	33.7	4.2	75	1782	2293	
	18	35.1	2.2	75	1932	2504	
	17.9	35	2.6	73	1863	2586	
	17.8	33.7	3.9	65	2115	1347	
	17.7	35	3.4	65	2127	1454	

- 対象となる範囲を選び、「標準」書式を指定する
- 桁区切りのカンマが取れていること、小数部分がきちんと表示されていることを確認
- カンマが残る場合はその部分を選択し、明示的に「数値」を指定してください。

CSV UTF-8形式で保存

Excel ブック (*.xlsx)
Excel マクロ有効ブック (*.xlsm)
Excel バイナリ ブック (*.xlsb)
Excel 97-2003 ブック (*.xls)
CSV UTF-8 (コンマ区切り) (*.csv)
XML データ (*.xml)
単一ファイル Web ページ (*.mht;*.mhtml)
Web ページ (*.htm;*.html)
Excel テンプレート (*.xltx)
Excel マクロ有効テンプレート (*.xltm)
Excel 97-2003 テンプレート (*.xlt)
テキスト (タブ区切り) (*.txt)
Unicode テキスト (*.txt)
XML スプレッドシート 2003 (*.xml)
Microsoft Excel 5.0/95 ブック (*.xls)
CSV (コンマ区切り) (*.csv)
テキスト (スペース区切り) (*.prn)
DIF (*.dif)
SYLK (*.slk)
Excel アドイン (*.xlam)
Excel 97-2003 アドイン (*.xla)
PDF (*.pdf)
XPS ドキュメント (*.xps)
Strict Open XML スプレッドシート (*.xlsx)
OpenDocument スプレッドシート (*.ods)

← CSV cps932形式
こちらは選ばない

← CSV UTF-8 を選択する

※ 古いOfficeでは UTF-8 が選べません。
その場合は CSV 形式で保存し、
データの読み込みのところで、
encoding='cps932' と書き換えてください。

空白データの除去

- 見えない空白が含まれている場合がある。
エラーの元となるので、削除しないといけない
- Unnamed や NaNが表示される場合は
その行 or セルを削除
- 表中に空白セルがある場合は念のため
削除するか0を入れる
- 検索置換機能で
空白を“”に置換
するのもよい
- 欠損のあるデータは
そもそも実験に
使わないのが無難

↑ ↓ ↶ ↷ ⚙ ↗ 🗑 ⋮

excel の csv ファイルの場合
の場合 (macOS, Linux はUnicode)

降水日数	雪日数	Unnamed: 10
4	145	0
4	132	5
0	122	3
1	125	6
5	134	17

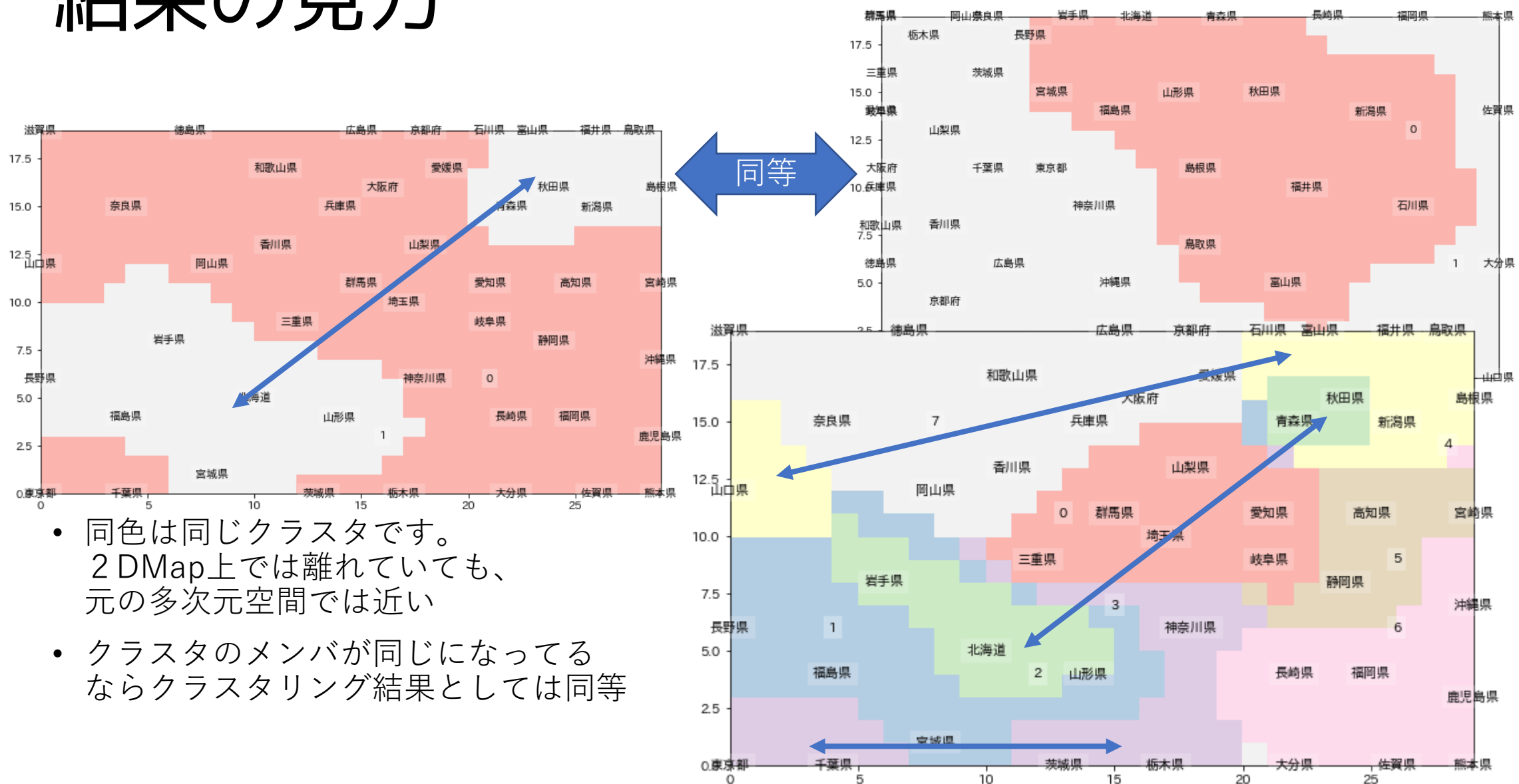
	I	J
	降水日数	雪日数
4	145	0
4	132	5
0	122	3
1	125	6
5	134	17
0	127	11
2	127	12
4	121	11
4	96	7
5	107	1
0	105	8
7	111	11
4	102	8
3	92	9
0	109	15
2	112	9
...

游ゴシック 11 A^ A^ % 0

B I ≡ ↶ ↷ A ↶ ↷ 0.00 0.00

✂ 切り取り(I)
📄 コピー(C)
📄 貼り付けのオプション:
📄 形式を選択して貼り付け(S)...
📄 挿入(I)
🗑 削除(D)
🗑 数式と値のクリア(N)
📄 セルの書式設定(E)...
📄 列の幅(W)...
📄 非表示(H)
📄 再表示(U)

結果の見方



結果のバリエーションについて

- 同じデータからできるクラスタリング結果は1通りではない。
- 重みの初期値に乱数を使ってためであるが、結果が無限にあるわけではない。

