# 星座グラフの変形による多変量データの可視化とクラスタ分析への応用

## A Modification of Consteration Graph and Its Application to Cluster Analysis

（2020年３月31日受理）

福森　護　藤原　美佳
Mamoru Fukumori　Mika Fujiwara

Key words：星座グラフ，クラスタ分析，スコッチウィスキー


## 要　旨

グラフ解析法の中で，クラスタ分析や判別分析と同様の目的で使用される星座グラフを変形させた新たな手法（以下，変形星座グラフ）を提案し，その有効性について，ウィスキーのフレーバーデータを用いて検証した．用いたデータは，12のフレーバーを0点から4点までの5段階で評価したものである．このデータに対して，K-means法により3つのクラスタに分類し，さらに変形星座グラフを適用することにより，その有効性を示した．さらに，従来の星座グラフと変形星座グラフとの比較を通して，それぞれの手法の利点や問題点，またクラスタ分析への応用の可能性などについて論じた．


## 1．は　じ　め　に

多変量データを２次元空間で可視化するためのグラフ表現法は，統計的データ解析における重要な手法の一つとして，社会科学をはじめ，様々な領域で活用されている．代表的なものとして，レーダーチャートや散布図などがあるが，それ以外にも独自のアイデアでさまざまな手法が提案されている．例えば，最大18個の変数を顔の部分に対応させて顔を描くことにより，顔の表情によって視覚的に多変量データの類似性の評価やグループの把握を可能にする顔型グラフ[1]や，多変量データの変数の値を，木の枝の傾き，葉の数，根の長さなどに対応させることにより，相関の度合いの視覚的な判定を可能にする木形グラフ[2]，変数の値を文字の大きさに対応させることにより，変数の性質を表す文字とその文字のサイズによって変数の特徴を可視化する文字グラフ[3]，また，重回帰分析の統計量をグラフ表現したボンサイグラフ[4]など数多くのグラフが提案され，利用されている．これらのグラフ表現法の中で，クラスタ分析や判別分析と同様の目的で利用されるグラフ表現法の一つとして，星座グラフ[5]がある．

星座グラフは，半円の中に，変数のベクトルを連結して最終点に星を描くことにより，データを可視化する手法である．変数の総合指標，平均得点，ばらつきの程度などを直感的に把握することが可能であり，ベクトルのパスを描くことにより変数の傾向や特徴を視覚的に把握できる．

星座グラフの描画の手順は次の通りである．

(1) 大きさ $n$ の $p$ 変量のデータ $x_{ij}$（$i=1, 2, \cdots, n$：$j=1, 2, \cdots, p$）が与えられたとき，変数を

$$\begin{cases} \theta_{ij}=f_j(x_{ij})\,;\, j=1,\ 2,\ \cdots,\ p,\ i=1,\ 2,\ \cdots,\ n, \\ \text{ただし，}\ 0 \leq \theta_{ij} \leq \pi \end{cases}$$

といった変換によってベクトルとして表す．

変換 $f_j$ としては，

$$f_j(x_{ij}) = \frac{x_{ij} - x_{jl}}{x_{ju} - x_{jl}}\pi\ ;\ x_{ju} = \max_{1\le i\le n} x_{ij},\ \ x_{jl} = \min_{1\le i\le n} x_{ij}$$

を考える．

(2) この$\theta_{ij}$を用いて，以下のようにベクトルの最終点の座標$(x, y)$を決定する．

$$(x, y) = \left(\sum_{j=1}^{p} w_j \cos\theta_{ij},\ \sum_{j=1}^{p} w_j \sin\theta_{ij}\right)$$

ここで，$w_j$は，最終点が半円の中に入るように，

$$\sum_{j=1}^{p} w_j = 1\ ;\ (w_j \ge 0,\ j = 1,\ 2,\ \cdots,\ p)$$

とする．

(3) 全てのサンプルに対して，同様にベクトルを連結し，最終点に星を描くことにより，図1に示すように，サンプルの数のパスと星が半円内に描かれる．

このように，星座グラフでは，半円の円周上に得点を割り当て，得点の方向にベクトルを連結させる方法である．ベクトルの長さについては，特にルールは定められておらず，一般には等ウェイトとして同じ長さのベクトルを連結させることが多い．

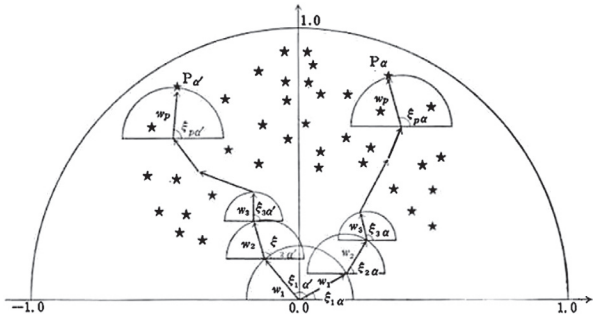


図1. 星座グラフの基本図（脇本他[5]より引用）

本論文では，星座グラフを応用したグラフとして，半円の円周上に変数を配置し，変数の方向にベクトルを連結させる新しい星座グラフ（変形星座グラフ）を提案する．また，提案手法の有効性を確認するために，ウィスキーのフレーバーデータに変形星座グラフを適用し，従来の星座グラフと比較することにより，その有効性や問題点について検討する．

## 2．変形星座グラフ

変形星座グラフは，半円の円周上に変数を布置し，ベクトルの長さに元データの得点を割り当ててベクトルを連結させ，最終点に星を描く方法である．描画の手順は次の通りである。

(1) 大きさ$n$の$p$変量のデータ$x_{ij}$($i = 1,\ 2,\ \cdots,\ n$：$j = 1,\ 2,\ \cdots,\ p$)が与えられたとき，次のような変換を考える．

$$\theta_j = \frac{j-1}{p-1}\pi\ ;(j = 1,\ 2,\ \cdots,\ p)$$

(2) この$\theta_j$を用いて，以下のようにベクトルの最終点の座標$(x, y)$を決定する．

$$(x, y) = \left(\sum_{j=1}^{p} s_j x_{ij} \cos\theta_j,\ \sum_{j=1}^{p} s_j x_{ij} \sin\theta_j\right)$$

ここで，$s$は変数の数によって定義され，以下の条件を満たしているものとする

$$s_j = \frac{1}{\max\limits_{1\le i\le j} x_{ij}}$$

$$x_{ij} \ge 0$$

(3) 全てのサンプルに対して，同様にベクトルを連結し，最終点に星を描くことにより，サンプルの数のパスと星が半円内に描かれる．

このように，変形星座グラフは，円周上に配置された変数の方向と各変数の得点を長さとする$p$個のベクトルを右または左から順に連結して，最終点に星を描く．同様に全てのサンプルについてベクトルのパスと最終点の星を描くと流れ星のような星座グラフができあがる．

## 3. 方　　法

### 3.1　使用データ

David Wishard[6] に記載されているスコッチウイスキーのテイスティングスコアを用いた．本スコアは，

フレーバーに関する12項目(Body, Sweetness, Smoky, Medicinal, Tobacco, Honey, Spicy, Winey, Nutty, Malty, Fruity, Floral)に対して，0点～4点の5段階評価となっている．記載されている100銘柄から，本研究では，20銘柄を選んで使用した．20銘柄の内訳は，Smokyが高得点である5銘柄，特に強い個性を持たずバランスよくフレーバーが含まれている9銘柄，Wineyが比較的高得点である6銘柄とした．

### 3.2 分析方法

まず，フレーバーの12項目に対して，k-means法によりクラスタに分類した．K-means法は各クラスタの中心と各データとの距離が最小になるように，データのクラスタへの割り当てを繰り返す手法である．そのアルゴリズムは以下の通りである．

1. 観測したデータ $x_i$($i$=1，2，…，$n$)にランダムにクラスタ $C_k$($k$=1，2，…，$K$)を割り当てる．
2. 各クラスタの中心 $x_k$を割り当てられたデータから求める．
3. 各 $x_i$と各 $x_k$との距離を求め，最も近いクラスタに $x_i$を割り当て直す．
4. 全てのデータに割り当てられたクラスタに変化がなければ終了し，変化があれば2に戻り，手順を繰り返す．

本研究では，クラスタの解釈可能性から3クラスタ解が妥当であると判断した．これらのクラスタの結果から得られたクラスタの特徴について調べるために，星座グラフ及び変形型星座グラフにより可視化を行った．

## 4．結　　果

### 4.1 変形星座グラフによる可視化

それぞれのクラスタの特徴を分析するために，変形星座グラフを適用し，図2に示すように，データの可視化を行った．なお，変数の並び順は，David Wishard[6]の記載順と同じとした．また，変数の配置の間隔は等間隔とした．

図2は，20銘柄それぞれのデータに対して，右側の変数から順にベクトルを連結してパスを描き，最終点に星を描いたものである．この図において，各変数と同一方向のベクトルの長さ及び最終点の位置によって，各クラスタの詳細な解釈が可能になる．図2から，20銘柄が3つのクラスタに分類されていることが読み取れる．また，パスの形状から，それぞれのクラスタの特徴の解釈が可能になる．

まず，図3は，クラスタ1の銘柄を濃い線で示したものである．クラスタ1の銘柄については，その多くが，Smokyの変数に対するベクトルが最も長くなっていることがわかる．さらに,Bodyも全体的にベクトルが長くなっており，一方で，SweetnessやFloralといった華やか系のフレーバーのベクトルが短いことが示されている．これより，クラスタ1はフルボディ&スモーキーフレーバーのクラスタと解釈される．

次に，図4は，クラスタ2の銘柄を同様に示したものである．クラスタ2の銘柄については，ベクトルの長さが，比較的SweetnessおよびFloralが長い銘柄が含まれているものの全体的に大きく長さが異なっていないことが特徴である．これより，クラスタ2はライトボディ&華やか系バランス型フレーバーのクラスタと解釈される．

最後に，図5は，クラスタ3の銘柄を同様に示したものである．クラスタ3の銘柄は，Sweetness, Honey, Winey, Fruityのベクトルが長い銘柄が含まれていることが特徴である．これより，クラスタ3は，ミドルボディ&華やか系フレーバーのクラスタと解釈される．

このように，クラスタ分析と変形型星座グラフを併用することにより，クラスタの特徴を詳細に分析できる利点がある．なお，本研究では，変数の並び順は，右から順に，引用文献の記載順に等間隔に配置した．

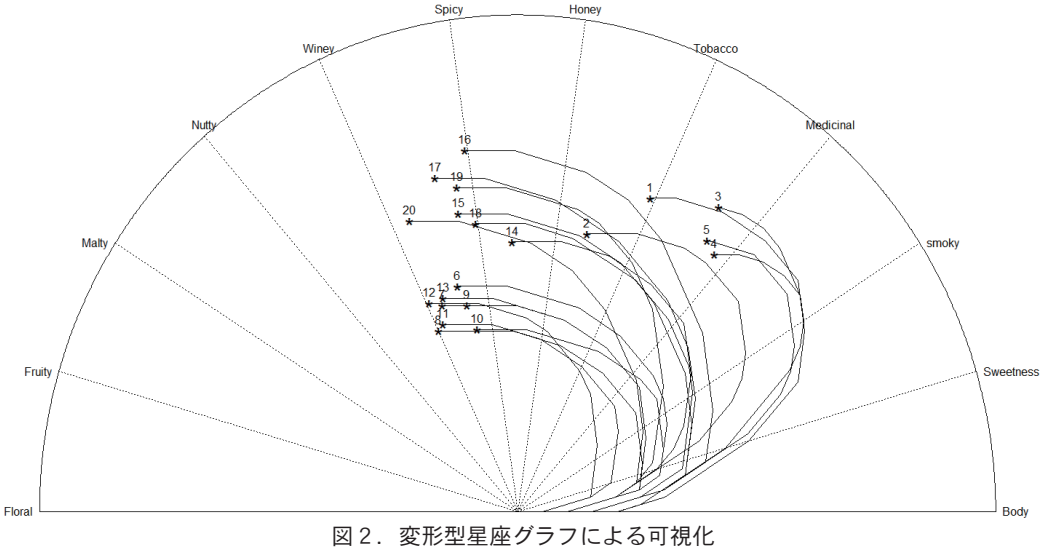


図２．変形型星座グラフによる可視化

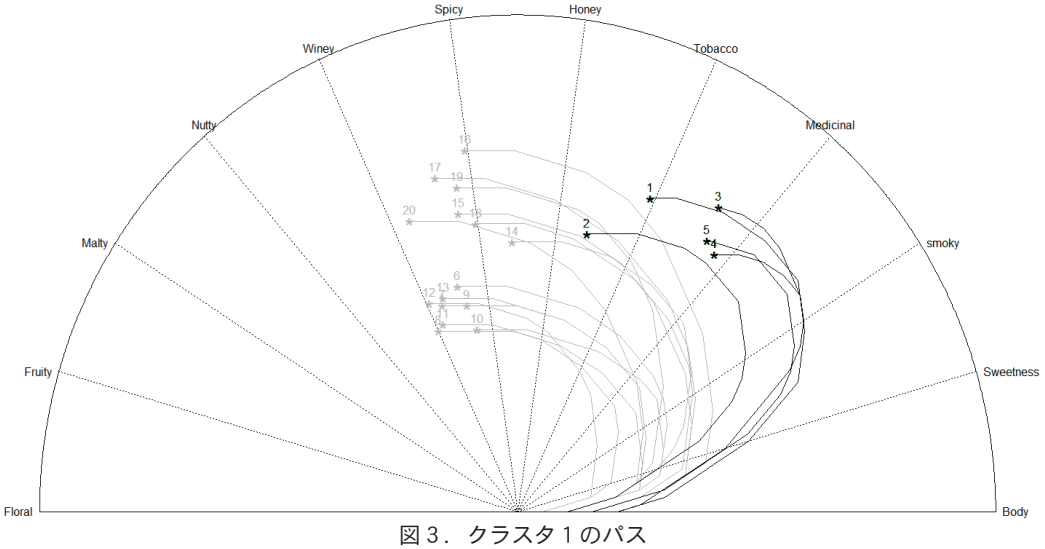


図３．クラスタ１のパス

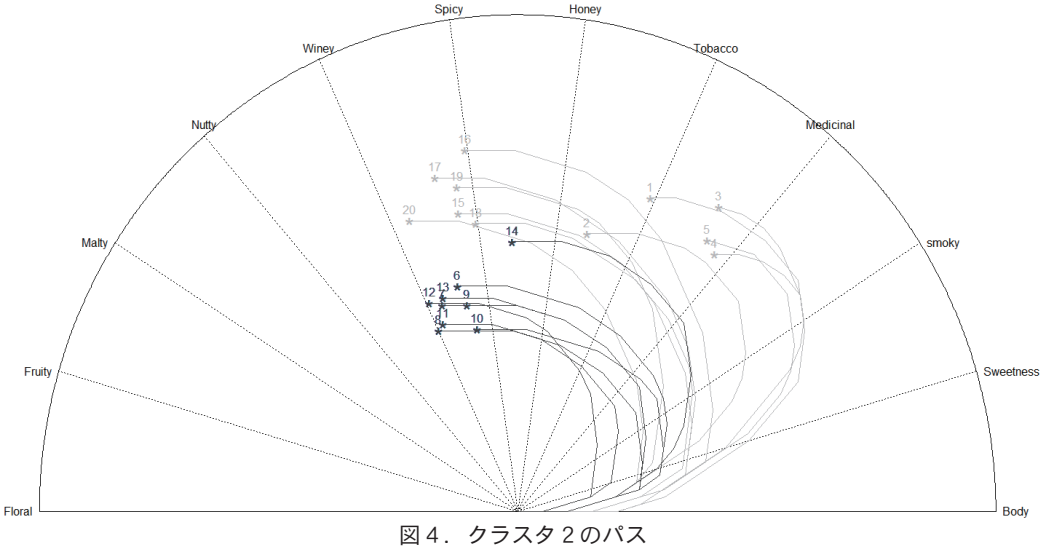


図４．クラスタ２のパス

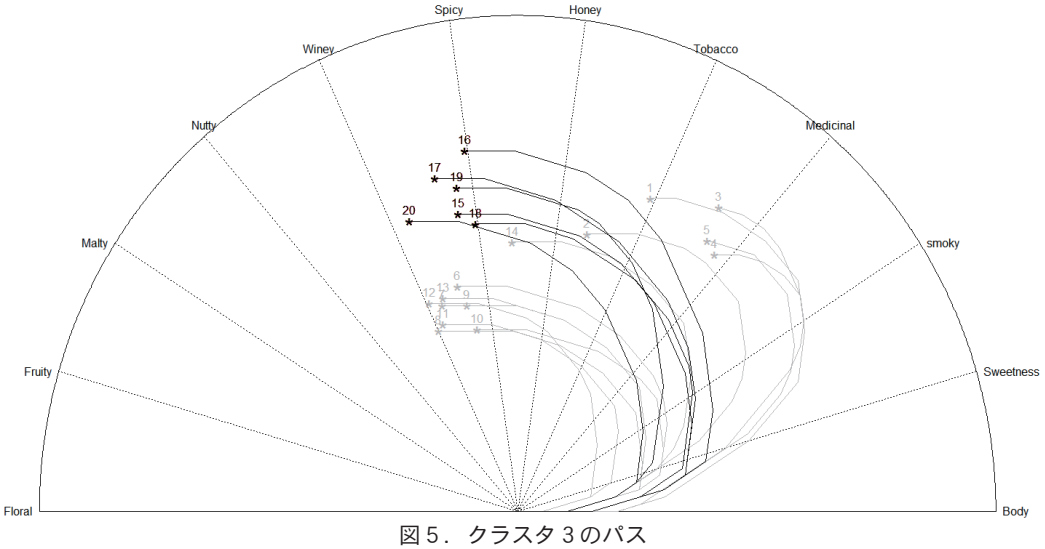


図５．クラスタ３のパス

### 4. 2　星座グラフによる可視化

比較を行うため，従来の星座グラフによる可視化を行った.星座グラフでは,5段階の得点を円周上に付置し,12の変数に対して得点の方向にベクトルを連結した．なお，ウェイトは等ウェイトとして描いた．また，変数の並び順は引用文献の記載順とした.

図6は，図2と比較して，3つのクラスタの分類を視覚的に把握することが難しい．また，視覚的に蒸留所の特徴を解釈することも容易ではない.

まず，図7は，クラスタ1の銘柄を濃い線で示したものである.

クラスタ1の蒸留所は，他の蒸留所よりも中心との距離が近い位置に最終点の星が付置している場合が多い.つまり，変数間の得点のばらつきが大きいことが示されており，特に強いフレーバーと弱いフレーバーがあることがわかる.従って,他のクラスタよりも個性的なフレーバーを持つ蒸留所が多いと解釈される.

図8は，クラスタ2の蒸留所を同様に示したものである.クラスタ2は，比較的右寄りに最終点の星が付置しているため,全体的にフレーバーの得点が低いことが読み取れる.

最後に，図9はクラスタ3の銘柄を示したものである.クラスタ3の銘柄は，他のクラスタよりも右に付置している銘柄が多いため，スコアの平均は最も高いクラスタであると解釈される.

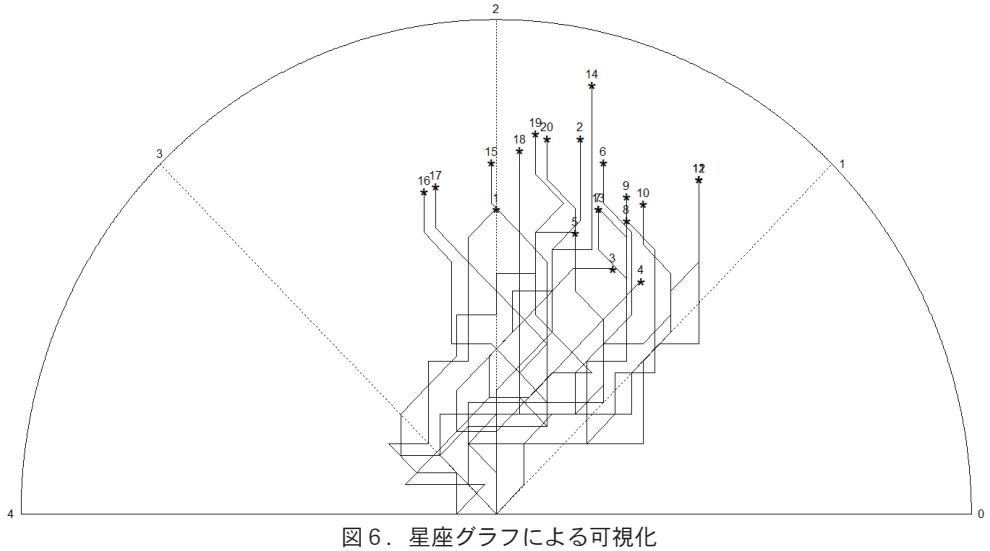


図６．星座グラフによる可視化

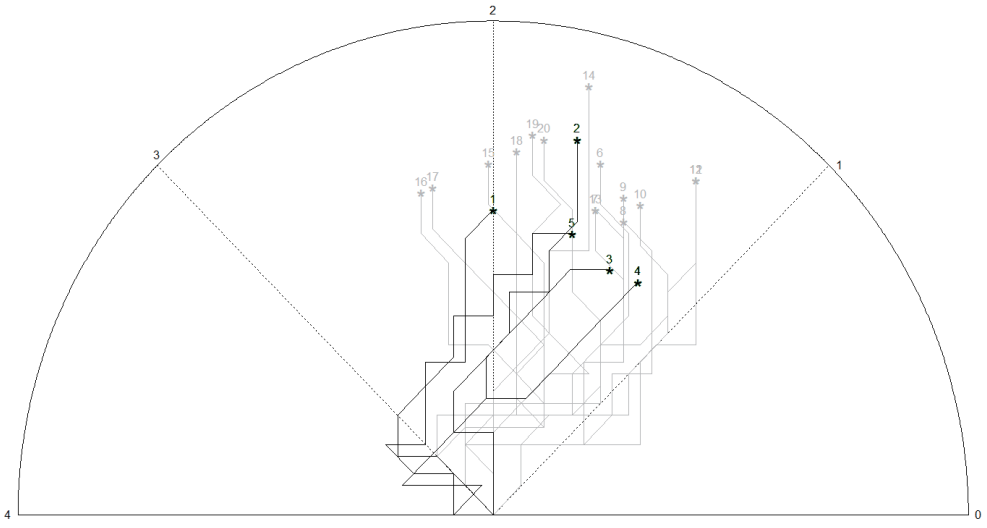

図 7．クラスタ 1 のパス

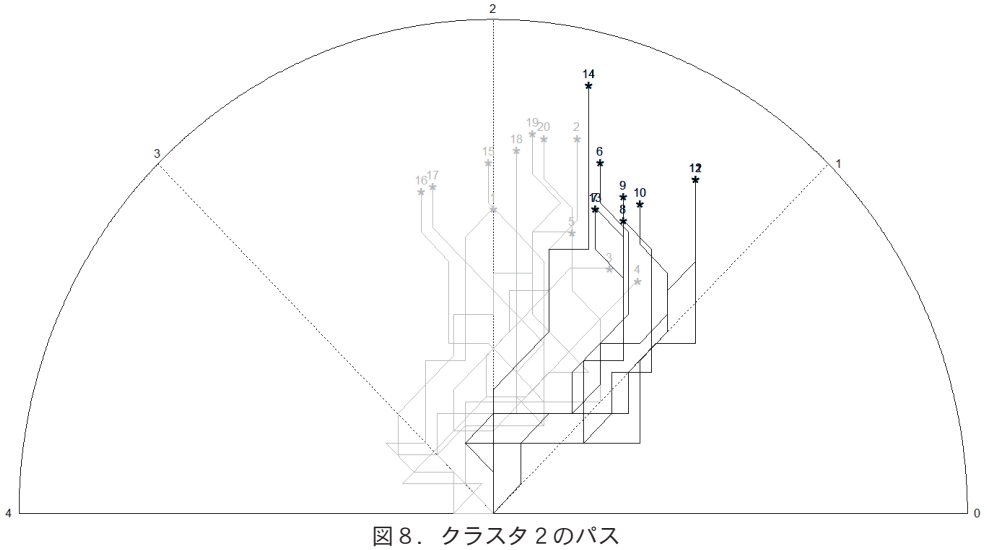

図 8．クラスタ 2 のパス

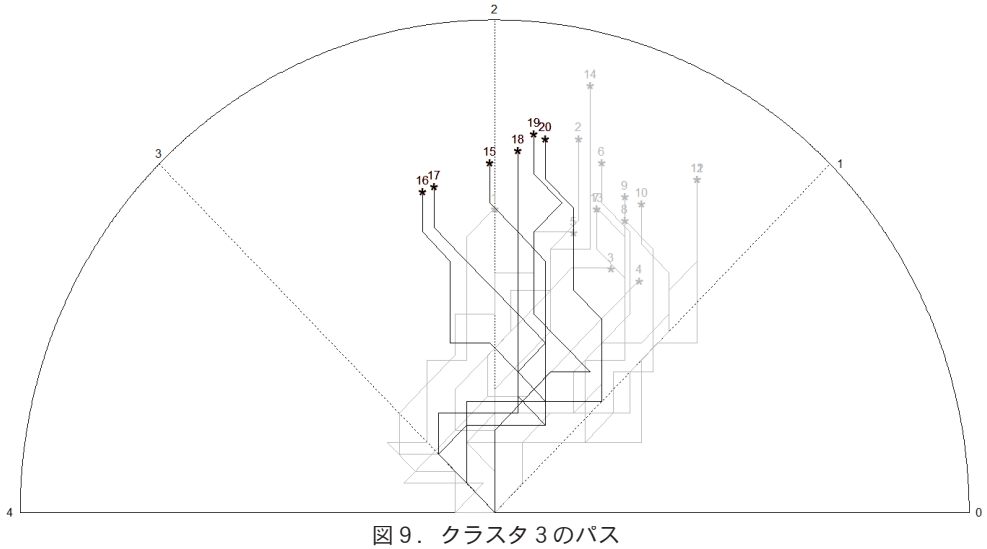

図 9．クラスタ 3 のパス

## 5. ま と め

以上，見てきたように，星座グラフでは，各銘柄のフレーバーのばらつきや合計・平均については直観的に把握できるという利点があるものの，フレーバーの特徴について視覚的に把握することは難しい．一方，変形星座グラフでは，各銘柄のフレーバーの特徴については視覚的に把握しやすいが，合計や平均，得点のばらつきについては把握しにくい．使用目的に応じて，使い分ける必要があるが，本研究のようなフレーバーによるウィスキーの分類及び各クラスターのフレーバーの特徴の分析という目的に対しては，銘柄のフレーバーの特徴が把握しやすい拡張型星座グラフが優れていると言える．

図2において，クラスタの分類が視覚的に示されているが，福森・田中[11]が指摘するように，グラフの知覚判断だけからデータの分類を行うことは望ましいことではなく，おおざっぱな分類においては効果を持つ場合があるものの詳細なデータの識別については限界がある．従って，変形星座グラフをk-means法の代わりとして利用することは望ましくない．本グラフは，あくまでクラスタ分析を相補する目的で併用した場合の有効性が高いと考えることができる．

変形星座グラフは，変数の並び順の課題が残る．何らかの指標に基づいて変数の並び順を決めることにより，パスの形状からおおまかな傾向と詳細な特徴の両方を読み取ることが可能になる．今後，さらなる改良が必要となることが示されたが，グラフ解析法は，用途に応じて柔軟に改良することが可能であるという利点がある．利用目的に応じた改良を加えることにより，さらに有効性な解析手法としての活用が期待される．

## 参 考 文 献

1) Chernoff, H. : The use of faces to represent points in k-dimensional space graphically, Journal of the American Statistical Association, 68, pp. 361-368, 1973.
2) Wakimoto, K. : Tree graph method for visual representation of mlti-dimensional data, Journal of the Japan Statistical Society, 7, pp. 27-34, 1977.
3) 平井安久，福森護，脇本和昌:. 多変量データの漢字グラフ表現とクラスタリングへの応用，計算機統計学，1(1)，pp. 11-21，1988.
4) 石村友二郎:重回帰分析のグラフ表現法，計算機統計学，26(2)，pp. 93-103，2013.
5) Wakimoto , K. & Taguril, M. Constellation graphical method for representing multi dimensional data, Annals of the Institute of Statistical Mathematics, 30, Part A, pp. 77-84, 1978.
6) Wishart, P. : Whisky Classified, PAVILION, pp. 36-41, 2018.
7) 福森護・田中豊:認知的観点による多変量グラフの評価-顔型グラフ，レーダーチャート，文字グラフの比較-，計算機統計学，7(1)，pp. 37-45，1994.