パーソナルコンピュータによる
ビッグデータのクラスター解析の妥当性

$PM_{2.5}$単一微粒子質量スペクトルに対して

森田　大智

1. 研究背景

1.1 $PM_{2.5}$ とは

空気中を浮遊する微小な液滴や粒子である大気微粒子の中でも粒径が $PM_{2.5}$ という。大気微粒子は工場の排煙や砂塵といった様々な発生源や発生プロセスが存在する。また、様々な粒径の粒子も存在し、それによって化学組成も多様化している。

主な化学組成成分としては、有機炭素，硫酸，硝酸，アルミニウム，塩素などがある。

1.2 単一微粒子質量分析計

単一微粒子質量分析計を用いて $PM_{2.5}$ の化学組成を測定する。単一微粒子質量分析計は空気中の微粒子を一つ一つを連続的に質量分析できる点がメリットである。測定方法としては、開口部から微粒子が入るとまず微粒子をビーム化し、その後レーザーによって微粒子をイオン化し、正イオンと負イオンそれぞれに対し電場をかけその飛行時間によって、質量電荷比を測定します。

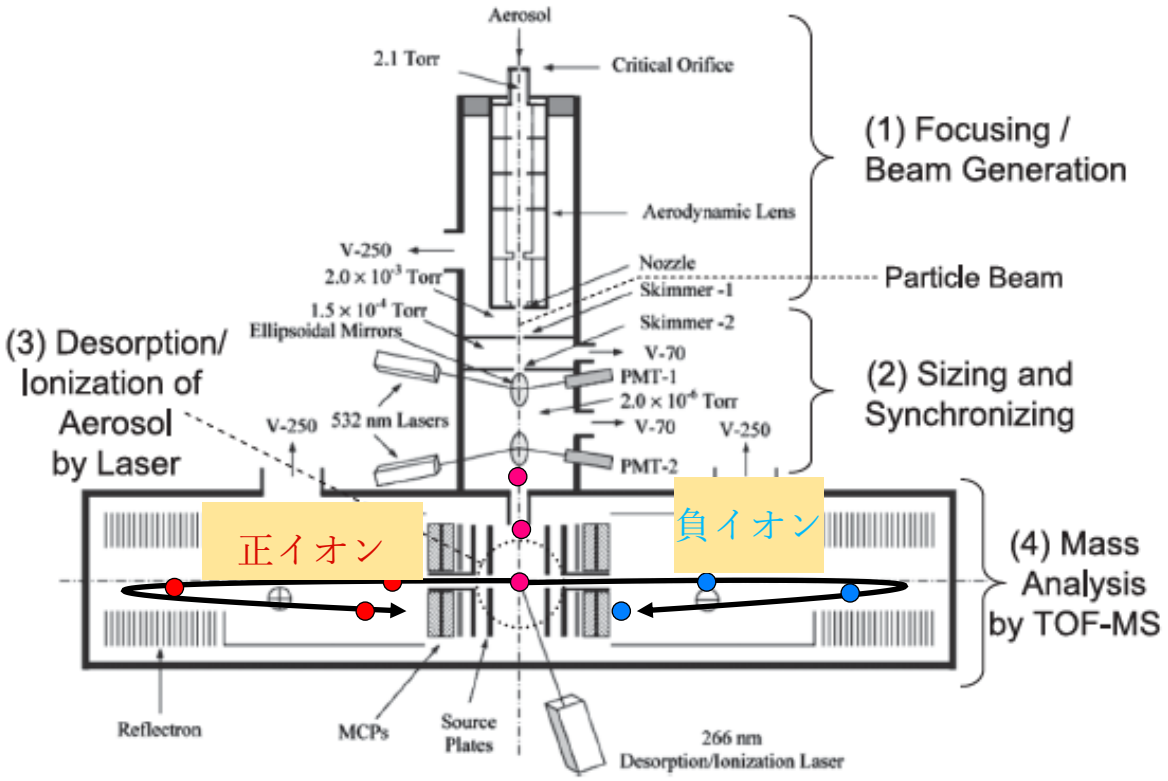

図 1　単一微粒子質量分析計

以下の図 2 が得られる質量スペクトルの一例である。

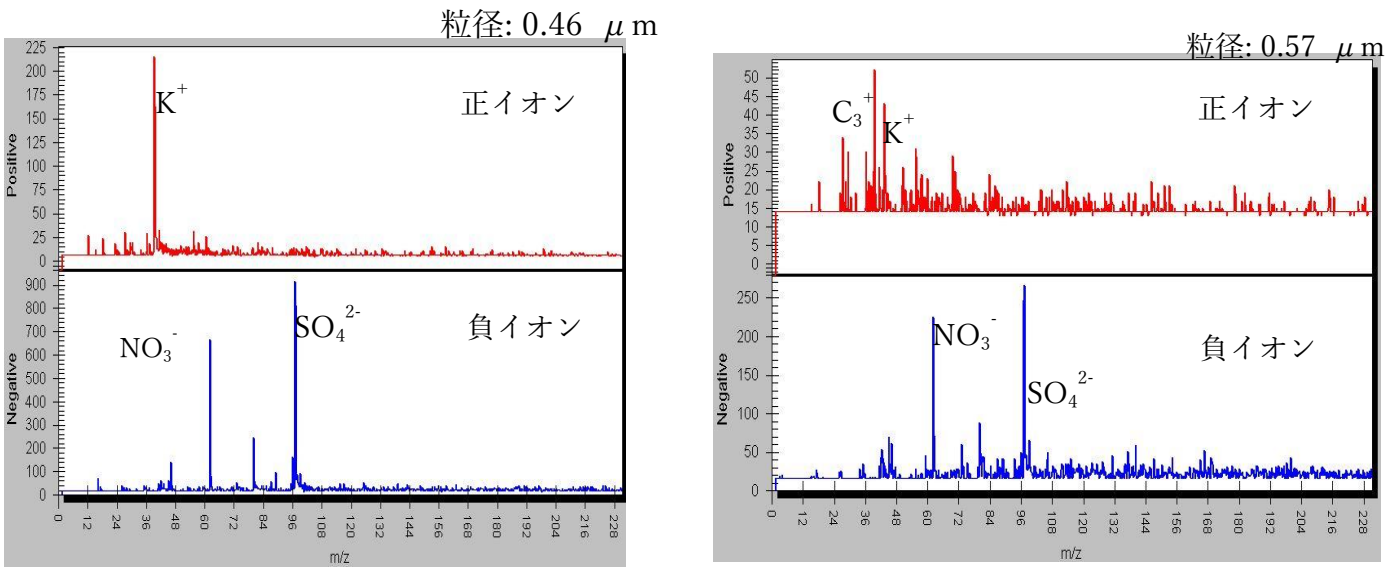

図 2　質量スペクトルの一例

この質量スペクトルの横軸は質量を表し、縦軸はそのイオン強度を表している。

1.3　今回用いるデータについて

今回データ解析するにあたって用いたデータは、単一微粒子質量分析計を用いて 2015 年 9 月 27 日から 10 月 18 日にかけて、中国北京市にある清華大学で行われた大気エアロゾルの観測によって得られた $PM_{2.5}$ 微粒子のなかで 10 月 1 日から 10 月 18 日のデータを用いた。測定した $PM_{2.5}$ 微粒子の数は約 260.7 万個であった。

1.4　クラスタリング

クラスタリングでは分類対象の性質に基づいて、いくつかのグループに分類するデータ解析の手法である。

今回、クラスタリングの手法として Art2a クラスタリングを用いた。

Art2a クラスタリングはほかのクラスタリング手法と比べ、もともとのグループ数を指定せずクラスタリング解析できるため、今回の測定のようにどれぐらいの数のグループが生じるか分からない場合に適したクラスタリング手法である。

観測した大気エアロゾルは正イオン、負イオンそれぞれ 1~350 のイオン質量で測定でき、合計 700 次元のベクトルとして扱う。そのベクトルを規格化し、それぞれのベクトルで内積をとりその内積の値が閾値以上であれば、同じグループとして閾値以上の内積が存在しなければ新たなグループとして取り扱うというクラスタリング手法を用いた。(閾値 0.80)

2. 研究目的

2.1 問題点

本来、クラスタリングするにあたり 260.7 万個の大気エアロゾルをまとめてクラスタリングしたい。しかしながら、一般のコンピュータでは約 10 万個のクラスタリングが限界である。そこで今回は、260.7 万個の中から約 10 万個を選び出し、その 10 万個に対しクラスタリング解析を行う。そのグループに対して残りの 250.7 万個を当てはめていくことでグループ分けを行った。

2.2 研究目的

スーパーコンピュータを用いて 260.7 万個をまとめてクラスタリングした結果（全数クラスタリング）と 10 万選び出しクラスタリングした結果（部分クラスタリング）を比較し、部分クラスタリングがどの程度特徴を捉えきれているか、また部分クラスタリングの妥当性を検討する。

3. 研究結果

3.1 クラスター数

全数クラスタリングと部分クラスタリングによって得られたクラスター数を比較すると全数クラスタリングでは 28914, 部分クラスタリングでは 2652 あった。数字にして約 11 倍の差が生じてしまっていることが分かる。この違いはどうして生じてしまっているのかをそれぞれのクラスターの持っている質量スペクトルの類似性を比較し検討する。

3.2　クラスター同士の類似性比較

比較手段としてそれぞれのクラスターの内積を用いた。（識別しやすいようにクラスターに含まれる粒子数の多い順にクラスター番号1, 2, 3,…とした。）例えば、図３のように部分クラスタリング１と全数クラスタリング 1~28914 の内積をそれぞれ計算する。 その後部分クラスタリング 2~2652 においても同様に内積を計算する。

図３クラスター数の比較方法

また、その結果が見やすいように図４のようにイメージ図にした。
内積の値が１に近いほど（類似性が高いほど）イメージプロットは赤色になり
内積の値が０に近いほど（類似背が低いほど）イメージプロットは青色になる。

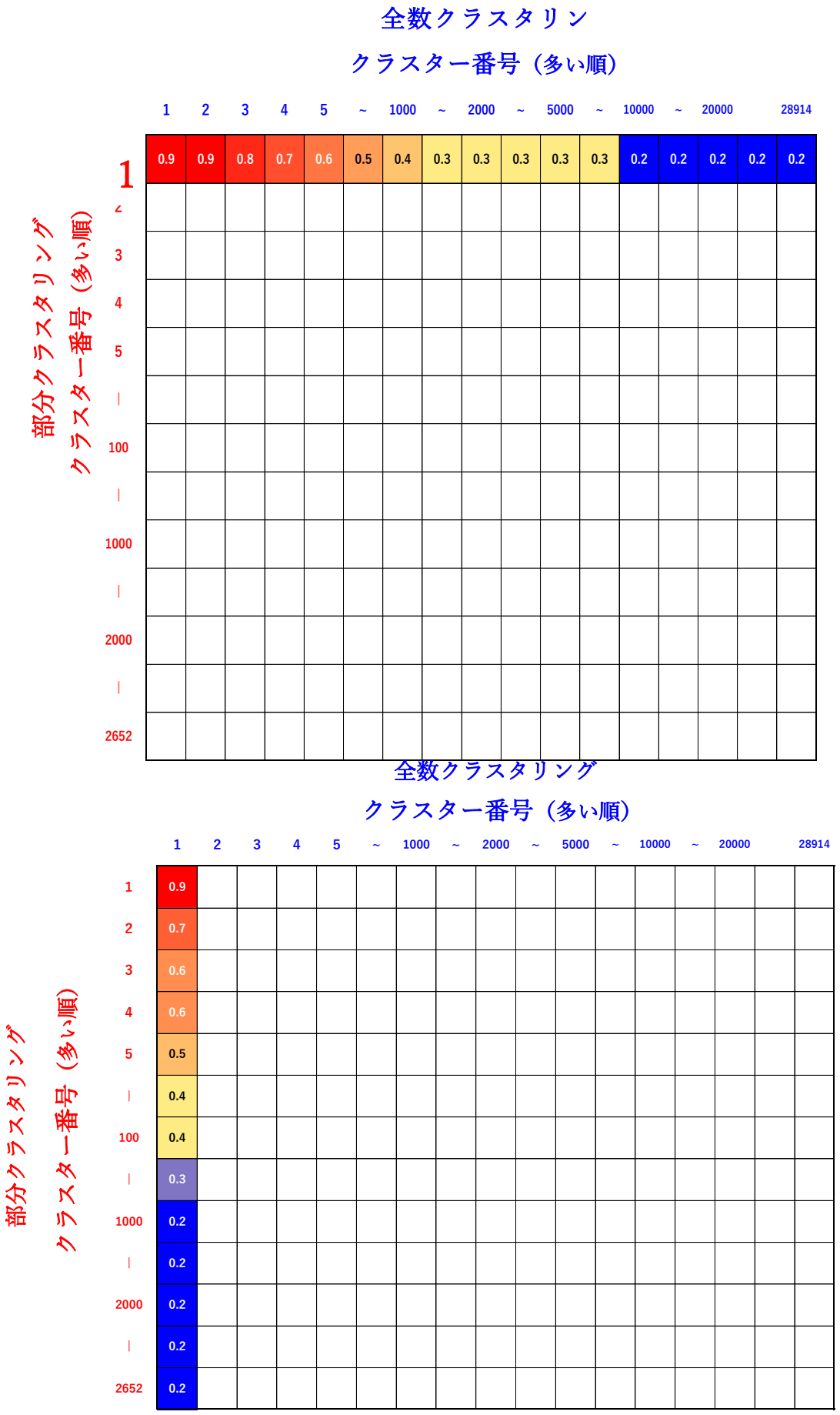

図４　類似性のイメージ図

実際にクラスターの質量スペクトル同士で内積をとりイメージプロットしたものが次の図５である。

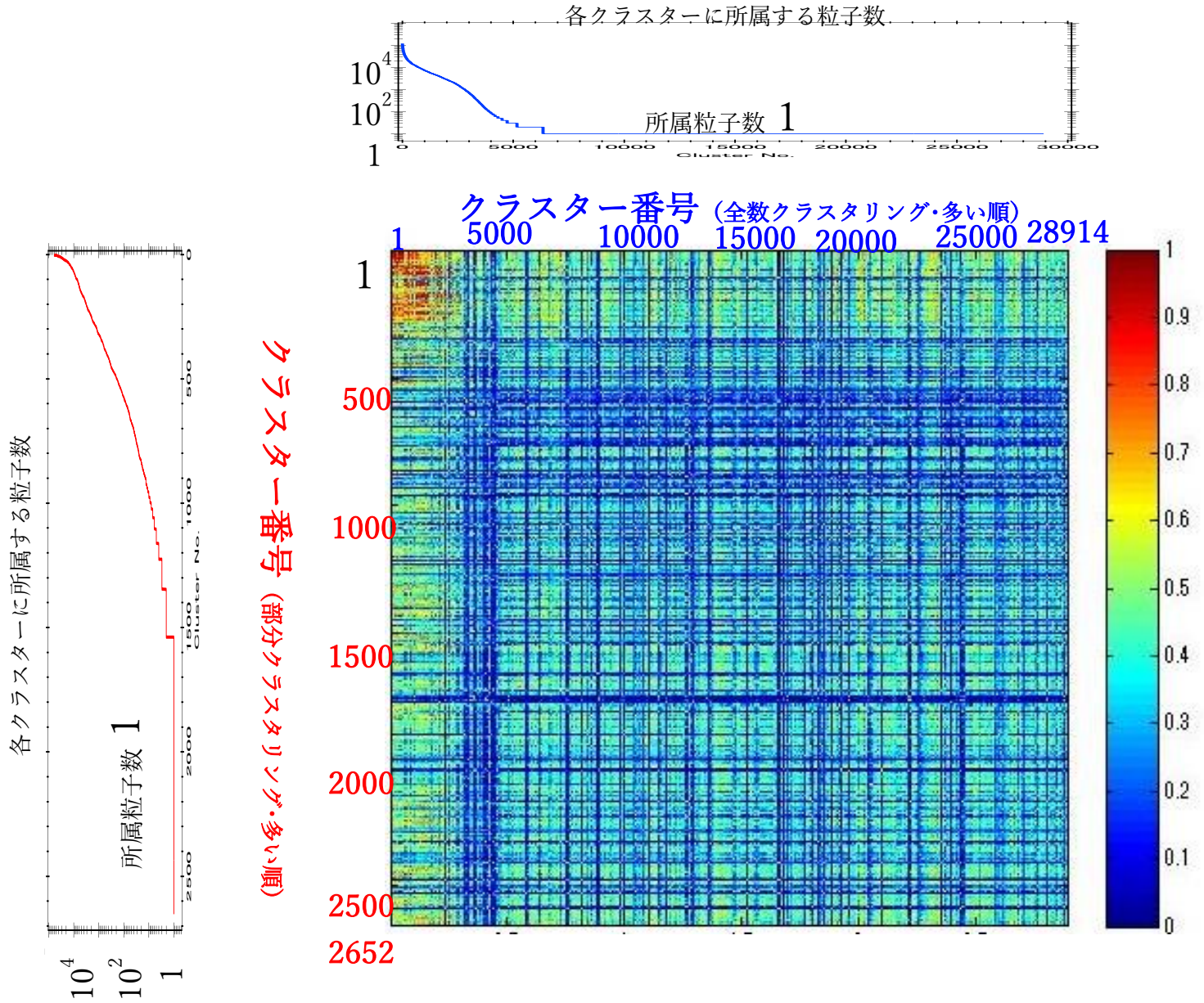

図５　クラスター同士の類似性のイメージプロット

イメージプロットの左側と上側のグラフはそれぞれ部分クラスタリングと全数クラスタリングの各クラスターに所属する粒子数を示すグラフである。そのグラフからもわかる通り所属する粒子数が 1 個のクラスターが大半を占めていることが分かる。所属する粒子数が１個のクラスターはノイズであるため、所属する粒子数の多いクラスターに注目する。次の図６が所属する粒子数が多いクラスターに注目したイメージプロットである。

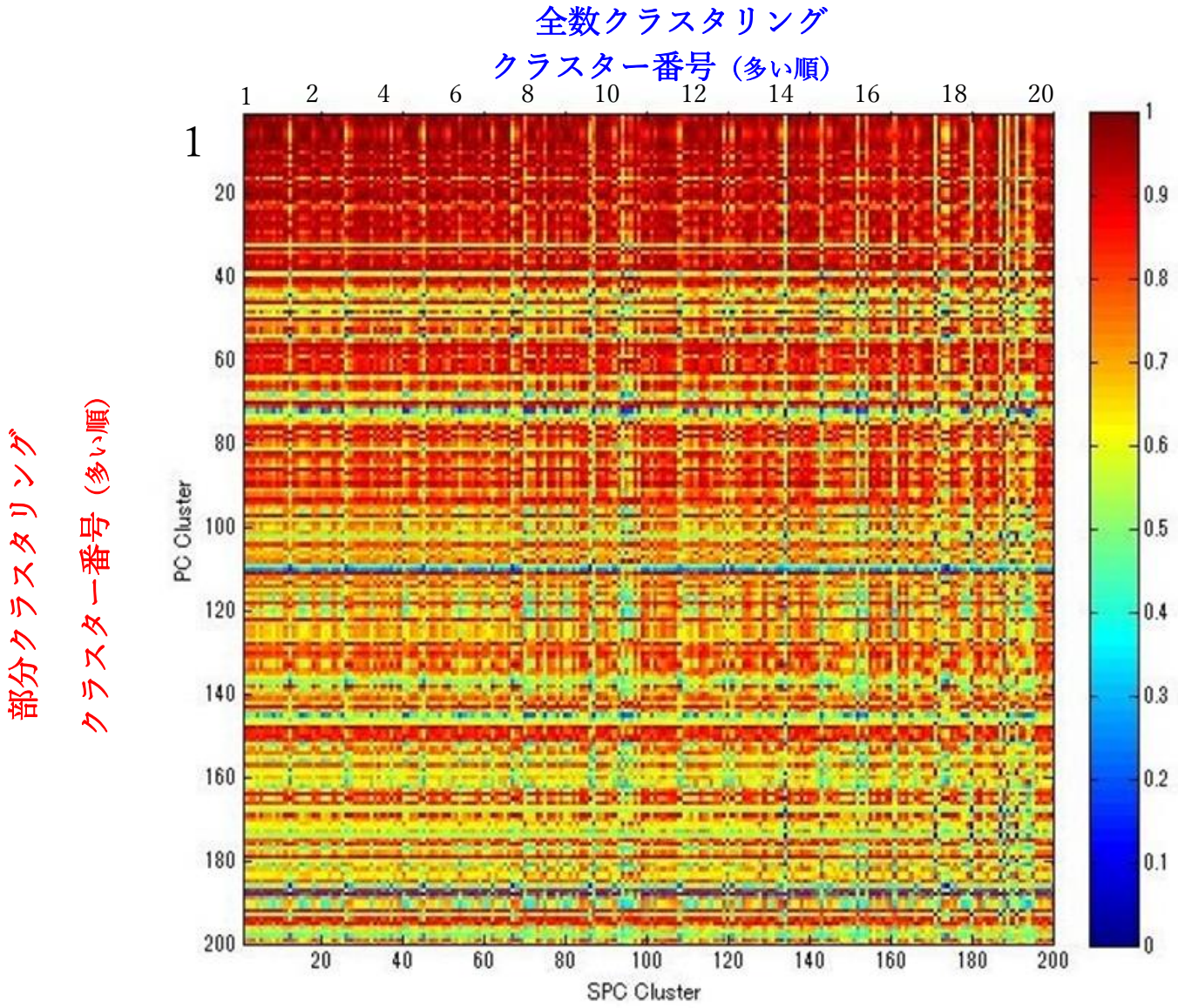

図 6　上位クラスターのイメージプロット

部分クラスタリングのクラスター40 番目と全数クラスタリングの 200 番目まで類似性の高いことを示す赤いエリアが広がっていることが分かる。実際は、部分クラスタリング、全数クラスタリング完璧なクラスタリングをしていれば対称になることが期待される。しかし、全数クラスタリングの方が多いクラスターに分類されてしまっているのが分かる。その原因を考察するため部分クラスタリング 1 番目と全数クラスタリング上位 10 クラスターの質量スペクトルを比較する。その質量スペクトルが次の図 7 である。

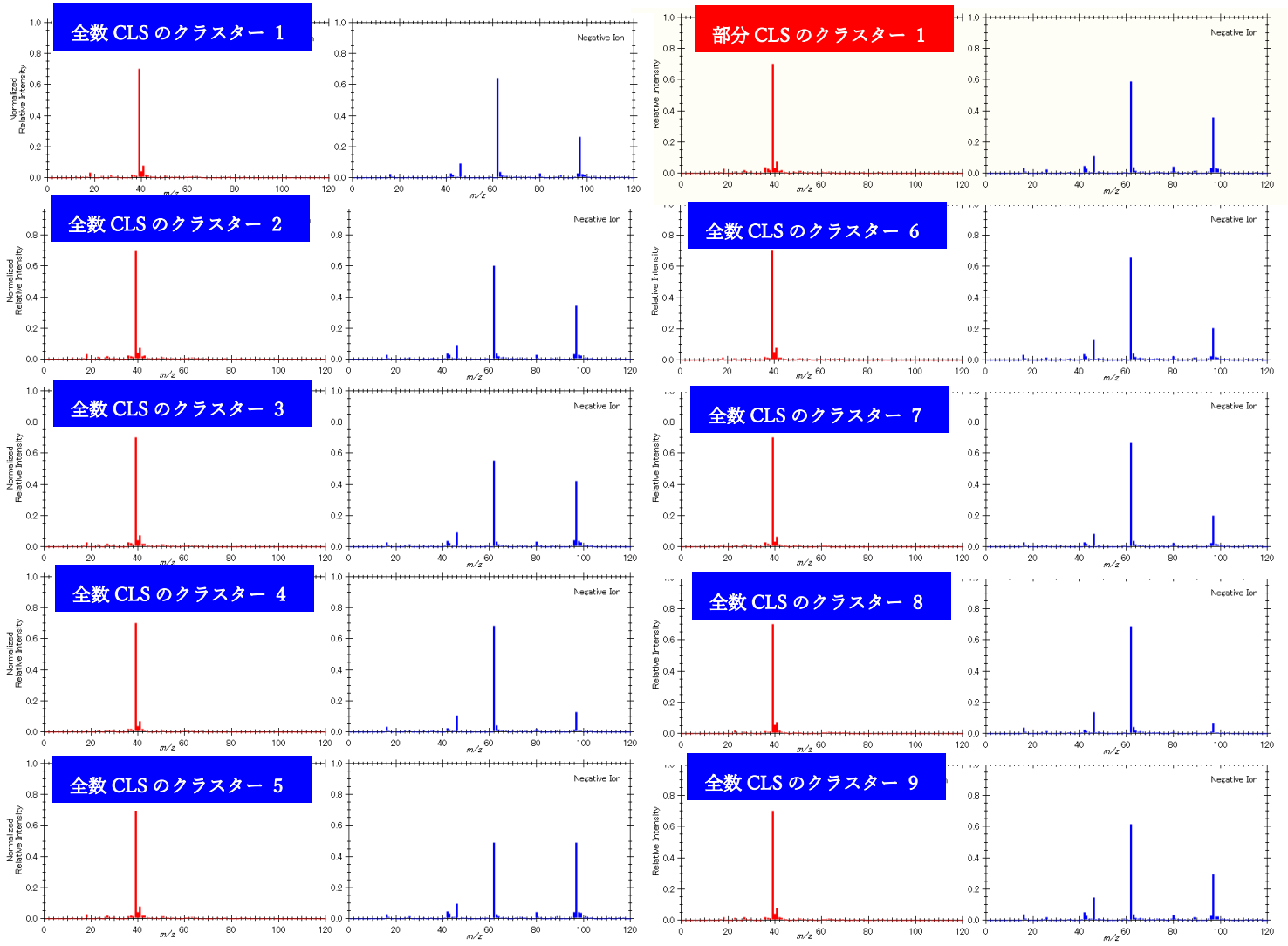

図 7　質量スペクトル

（部分クラスタリングクラスター1 番目と全数クラスタリング上位 9 クラスタ

それぞれの質量スペクトルを比較するとイオン質量のピークの質量は同じであるがそれぞれ微妙にイオン強度が異なってしまっていることが分かる。つまり、全数クラスタリングでは部分クラスタリングでは区別していない微妙な違いまでも区別してしまっていることが分かる。これは数学的な観点からすると全数クラスタリングは正しく区別できているが、大気化学的な観点から化学組成を見るうえで大きな違いはないといえる。

3. まとめ

上位クラスターでは部分クラスタリング、全数クラスタリングも同様のク

ラスターに区別されていることが期待されていたが、全数クラスタリングでは細かく区別されていることが分かった。しかし、大気化学的な観点では細分されすぎている。部分クラスタリングでも上位クラスターの大きな特徴はとらえられていることが分かった。

→つまり、部分クラスタリングでは主要なクラスターの大まかな化学組成を知る上では十分に活用できることが分かった。