

PM_{2.5} 化学組成解析のための個別粒子質量スペクトルのビッグデータ分析: 北京市での PM_{2.5} 観測を例に

Big Data Analysis of Single Particle Mass Spectrum Data for Chemical Characterization of PM_{2.5}: An Example from the Analysis of Beijing PM_{2.5} Observation

古谷浩志^{*1,*2} NAT Pavasant^{*3} 福井健一^{*4} 前田幸輝^{*2} 豊田岐聡^{*2} 紀本岳志^{*5}
FURUTANI Hiroshi FUKUI Kenichi MAEDA Kouki TOYODA Michisato KIMOTO Takashi
MA Tao^{*6} DUAN Fengkui^{*6} MA Yongliang^{*6} HE Kebin^{*6}

^{*1} 大阪大学 科学機器リノベーション・工作支援センター

Support Center for Scientific Instrument Renovation and Custom Fabrication, Osaka University

^{*2} 大阪大学 理学研究科附属基礎理学プロジェクト研究センター

Project Research Center for Fundamental Sciences, Graduate School of Science, Osaka University

^{*3} チュラロンコン大学計算機工学科

Chulalongkorn University

^{*4} 大阪大学産業科学研究所

Institute of Scientific and Industrial Research, Osaka University

^{*5} 紀本電子工業株式会社

KIMOTO Electric Co. Ltd.

^{*6} 清華大学 環境学院

School of Environment, Tsinghua University

PM_{2.5} is fine liquid or solid particulate matter suspended in the atmosphere, which affects not only human health but also earth's environments (e.g., acid rain and global warming). In this paper, we will present an application of big data analysis to the single particle mass spectrum data obtained from individual several to tens of million PM_{2.5} particles in Beijing as an example of the application of the big data analysis for PM_{2.5} chemical characterizations, such as chemical composition and mixing, size distributions, and their temporal variation and emission sources.

1. はじめに: PM_{2.5} と大気化学について

大気中には、微小な固体粒子や液滴など、一般には“塵”と呼ばれる粒子状の物質が浮遊している。これらの浮遊物質はエアロゾルと呼ばれ、その中でも粒径が 2.5 μm「以下」の微小なものは特に PM_{2.5} と呼ばれている。近年の経済発展にともなう、世界各地で PM_{2.5} 濃度の増加による甚大な大気汚染が発生しており(例えば北京市、ニューデリー市など)、人への健康影響が懸念されている。日本においても、越境大気汚染などの影響により PM_{2.5} 濃度が一時的に増加する状況が見られ、社会的な関心が非常に高まっている。このような健康影響だけでなく、PM_{2.5} は地球温暖化問題にも深く関係している。というのも、PM_{2.5} を含む大気エアロゾルは、太陽光を直接・間接的に散乱・反射することで地表への日射量を減少させ、温室効果ガスによる温室効果を緩和する働き(冷却効果)を持っているためである。PM_{2.5} は人間だけでなく、人間の生存する環境に対しても様々な影響を与えているといえる。

一括りに「PM_{2.5}」とされることが多い大気エアロゾル粒子であるが、その物理・化学的な特性は非常に複雑かつ多様である。粒径は 1 nm から数 10 μm と 5 桁と広く、またその化学組成も無機物、金属類、有機物と複雑かつ多様である。物質によって毒性が異なるように、PM_{2.5} の健康・環境影響の度合いも、PM_{2.5} 粒子の化学組成やそれらの混合具合、粒径や粒子数によって大きく異なる。このように PM_{2.5} 粒子の化学組成、濃度、それらの粒径分布などを観測し、状況を明らかにすることは PM_{2.5} 研究の出発点であるといえる。

しかし、現状把握から一歩進み、どのように PM_{2.5} を減らすかという問題解決策を考えようとするならば、これらの基本特性を知るだけでは不十分である。なぜなら PM_{2.5} 粒子には、図 1 に示すように様々な発生源から微粒子として直接放出される成分だけでなく、ガス状物質が大気中で化学変質することで二次的に粒子化する成分(二次粒子成分)とがあり、更に複雑なことに、2つ

の成分は互いに大気中で混合していくのである。

「PM_{2.5} が何からどれだけ大気中に放出され、また大気中でどのような化学機構でどれだけ生成されるか」といった「大気中の化学物質の挙動」を扱う学問分野は、大気化学(Atmospheric Chemistry)と呼ばれ、PM_{2.5} は大気化学研究の大きなトピックの1つである。

本発表では、先端 PM_{2.5} 化学組成測定装置の1つである単一微粒子質量分析計(ATOFMS)による大気観測によって、数 100 万 - 数 1000 万個の PM_{2.5} 粒子から得られる個別粒子質量スペクトル「ビッグデータ」を、どのようにデータ解析し、大気化学研究に活用しているかについて紹介する。そして現在のデータ解析法の課題とデータサイエンティストとの協働(協力)の必要性について述べる。

2. PM_{2.5} とビッグデータ解析

2.1 単一微粒子質量分析計で観測した数 100 万～数 1000 万個の個別粒子質量スペクトルを大気化学の観点から解析する

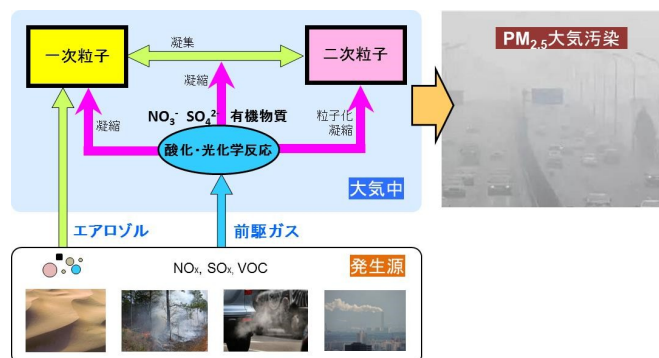


図 1 大気エアロゾルの放出・生成プロセスの概略図。発生源から微粒子として直接大気へ放出される一次粒子だけでなく、一次粒子と同時に放出される窒素酸化物(NO_x)、硫黄酸化物(SO_x)、揮発性有機ガス(VOC)などの微量ガスが、大気中での酸化・光化学反応によって化学変質し、二次的に硝酸塩(NO₃)、硫酸塩(SO₄²⁻)、有機物質(OC)などの粒子態物質となるプロセスの2つが並存し、互いに影響し合っている。

連絡先: 古谷 浩志 hiroshif@reno.osaka-u.ac.jp

筆者らは、「粒子1つ1つの粒径と化学組成を直接測定できる質量分析計」である単一微粒子質量分析計(ATOFMS)を用いて、大気中のPM_{2.5}粒子を連続観測し、その観測データを基にPM_{2.5}粒子の化学的性状やその粒径分布、PM_{2.5}発生源や生成機構を明らかにする研究を行っている。

単一微粒子質量分析計は、図2に示すように微粒子を個別に1つ1つずつ高強度のパルスレーザー光で蒸発・イオン化させ質量分析する先端エアロゾル化学分析装置である。測定された質量スペクトルは、各々の微粒子の化学組成に対応したイオンピークが含まれており、その質量スペクトルを解析することで各微粒子の詳細な化学組成を知ることができる(図3)。

単一微粒子質量分析計の特徴は、①有機物・無機物・金属類など多様な化学成分を一括して、しかも微粒子ごとに測定でき、また②その測定を連続してその場で高速に行えることである。不揮発性の金属・無機物質の中には、特定のPM_{2.5}粒子放出源の指標となる物質が多くあり、放出源の推定を行う上で非常に有用である。また放出源に特有の化学物質が、どのような二次物質と混合しているかという情報から、PM_{2.5}粒子が大気中でどのような化学変質を受けたのかを知ることができる。

従来のPM_{2.5}化学分析手法は、多くの場合に“PM_{2.5}サンプルを採取し、ラボに持ち帰り化学分析”というOff-Lineでの分析であり、化学物質の種類ごとに異なる分析装置が必要であるなど、多様な化学物質を同時に連続測定することは容易ではない。また、これらの化学分析は、多数のPM_{2.5}粒子の平均的描像を見るバルク分析でしかない。単一微粒子質量分析計では、PM_{2.5}粒子1つ1つに対する化学組成測定を行っているため、得られる情報の種類と量が従来法とは大きく異なっており、PM_{2.5}粒子に関する新しい知見を得るには好適であるといえる。

しかし、得られるデータ量は莫大となり、大気化学者にとってそのデータ解析は大きなハードルとなる。例えばPM_{2.5}濃度が

高い都市域などでは、1時間あたり約1万個の大気エアロゾルを測定でき、数週間から数ヶ月間の観測を行うと、数100万個から数1000万個という大量のPM_{2.5}粒子からの質量スペクトルが得られる。1個の質量スペクトルを10秒かけて目視で吟味すると、1万個の質量スペクトルの検討に必要な時間は10万秒(約28時間)。僅か1時間の大気観測で測定される質量スペクトルデータの解析だけで破綻してしまう。ビッグデータ解析が必要となる。さて、この単一微粒子ビッグデータをどのように大気化学の観点から解析していくか？

2.2 大気化学者が知りたいこと

まずPM_{2.5}粒子の化学的性状、粒径分布、それらの時間変動を知ることが大気化学の議論のスタートラインとなる。加えて単一微粒子質量分析計では、これらの情報を単一微粒子レベルで連続測定できる。この特徴を生かすために、単一微粒子質量分析計による観測データの解析では、①各々のPM_{2.5}粒子がどのような化学物質から成り、またどのような物質と混合しているのか(特定の混合パターンはないか？特定の化学組成を持つ粒子組成タイプに分類できないか？)、②それらの粒径分布や粒子数の時間変動はどうなっているか、③それらの時間変動が気象条件や他の原材料であるNO_x、SO_x、NH₃、有機ガス(VOCs)などの微量ガスとの時間変動と何らかの関係がないかなどを明らかにするアプローチを採ることが多い。

上記の①-③について大気化学研究の観点から吟味を行うためにも、まずは単一微粒子質量分析計で得られた質量スペクトル“ビッグデータ”を圧縮し、更にそれらを“大気化学データ”に変換しなくてはならない。

3. ART-2a を用いた個別粒子質量スペクトルビッグデータのクラスタリング解析

3.1 クラスタリング解析で粒子組成タイプに分類

まず最初に知りたいことは化学組成であり、加えて観測データを人が直接扱えるデータ量に圧縮するために、個別粒子質量スペクトルに対してクラスタリング法を適用し、類似した質量スペクトルすなわち類似した化学組成をもつPM_{2.5}粒子群(粒子組成タイプ)ごとに分類している。個別粒子質量スペクトルデータであるが、図3に示すように1つのPM_{2.5}粒子が、正イオンと負イオンの質量スペクトル(m/z 1 ~ 500, m/z -1 ~ -500)と粒径、測定時刻の4種類の情報を持つ。

クラスタリング解析のアルゴリズムとして、ART-2a (Adaptive resonance theory-2a)を用いている [Song 1999]。ART-2aの利点は、類似性判別の閾値であるVigilance Factor (1が完全に一致)と学習率であるLearning Rateを与えるだけで、先見情報なしに新しいクラスター(粒子組成タイプ)を追加しながらクラスタリングが行われる点である。この「先見情報なしに」の部分が、大気化学研究の立場から非常に重要である。というのも大気エアロゾルの組成は場所や時間などによって大きく変わるため、データ解析をしてみないと、どのような粒子組成タイプが何種類あるか分からないからである。

最もポピュラーなクラスタリング法であるK-Means法は、計算速度が高速であるためビッグデータ解析などに広く用いられている。しかし、「クラスタリング解析を行う前に」、最終的に何個のクラスターに集約させるか決めなくてはならない。適切なクラスター数の指定が、良いクラスタリングの結果を得るために不可欠であるが、複雑な自然相手では、事前に決定することは困難であり、ビッグデータに対しては尚更である。そのためアルゴリズムとしてART-2aを用いている。

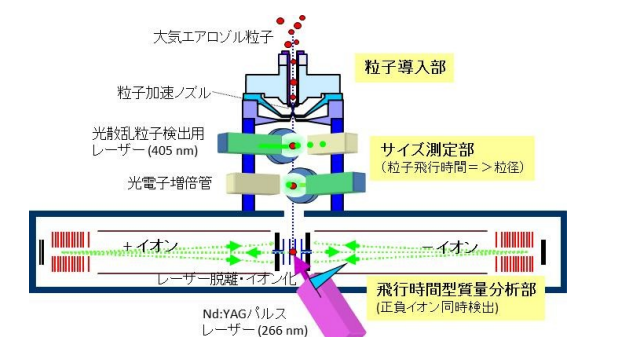


図2 微粒子1つ1つの粒径と化学組成をその場で連続的に測

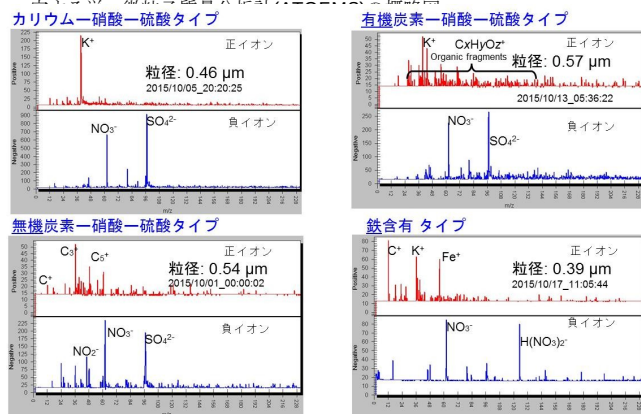


図3 単一微粒子質量分析計によって測定された、1個の微粒子からの質量スペクトルの例。質量スペクトルのイオンピークから、微粒子に含まれる化学組成が分かる。

3.2 ART-2a を用いたクラスタリング解析の実際と課題：北京市での大気観測データの解析を例に

図 4 は 2015 年に中国の北京市で清華大学の研究グループと共同で行った大気観測の様子である。室外から大気を吸引しその吸引大気中に含まれる $\text{PM}_{2.5}$ 粒子を、実験室内に設置した単一微粒子質量分析計 (ATOFMS) で連続測定した。2015 年 10 月期には 10 月 1 日から 10 月 18 日までの 18 日間の大気観測を行い、粒径 100 nm から約 2 μm までの大気エアロゾル 260.7 万個の化学組成と粒径を測定した。



図 4 清華大学 (北京市) での単一微粒子質量分析計 (ATOFMS) による大気観測 (左)。実験室からの眺め (右)。

測定した個別粒子質量スペクトルデータは、YAADA (Yet Another ATOFMS Data Analyzer) と呼ばれるオープンソースの解析ツールを用いて解析している。ART-2a を用いたクラスタリング解析もそのツールの 1 つとして提供されており、それを利用している。YAADA は代表的な数値解析ソフトウェアの 1 つである Matlab 上で動作するため、筆者らは一般的なパーソナルコンピュータを使ってデータ解析を行っている。

図 5 は、データ解析の流れである。まず 260.7 万個の $\text{PM}_{2.5}$ 粒子から約 10 万個を選択し、それに対して ART-2a クラスタリング解析を行い、種となる粒子組成タイプ (Seed クラスタ) を決定し、続いて残りの 250.7 万個をこの Seed クラスタにマッチングさせて、全粒子を分類している。本来ならば、260.7 万個の全粒子データに対して直接 ART-2a を行うべきであるが実現していない。またマッチングの過程では、ART-2a のように新たなクラスタを追加することはなく Seed クラスタは固定されたままである。260.7 万個の全粒子データに対して直接 ART-2a クラスタリングが適用できない理由であるが、それはメモリの制限のためである。ART-2a の計算速度を向上させるため、全粒子の質量スペクトルデータを行列データとしてメモリ上に置き、それに対して各粒子の質量スペクトルの類似性をチェックしている。この行

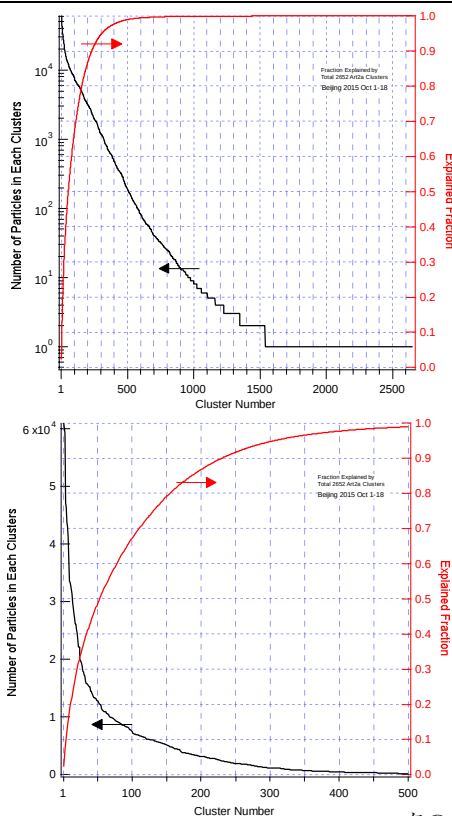


図 6 260.7 万個の個別粒子質量スペクトルデータをクラスタリングした結果。2653 個のクラスターに分類され、クラスターに属する粒子数が多い順番に示してある。赤線はそのクラスターまでの粒子で、全粒子のどの割合を説明出来るかを示している。ART-2a クラスタリングは Vigilance Factor 0.8 で Learning Rate 0.05、繰り返し 20 回の条件、マッチングの Vigilance Factor も 0.8 とし、類似性の判断には規格化した質量スペクトル同士の内積を用いている。

列計算を行うために必要なメモリが制約となり、8-16GB のメモリを装備した PC でも約 10 万個 (配列数 $10^3 \times 10^5$) が ART-2a クラスタリングの限界となっている。

10 万個の粒子データは 260.7 万個の粒子データの約 3.8% でしかない。もし 1000 ~ 2000 万個の粒子データの解析をしようとすれば、

わずかに全体の 0.5~1% でしかない。主要な粒子組成タイプの探索などの限られた解析用途には問題ないだろう。しかし、粒子数は少ないが特徴的な粒子組成タイプについての詳細解析などを行う場合には、全体のわずかに数% 以下の粒子群から得られた粒子組成タイプ (Seed クラスタ) の妥当性や代表性に疑問符が付く。大きな課題の 1 つである。

図 6 は、北京市で 2015 年 10 月期に観測した $\text{PM}_{2.5}$ 粒子の 260.7 万個に対して行ったクラスタリングの結果である。2653 種類のクラスターに分類されたが、その分布は典型的な Long Tail 型であることが分かる。上位 100 クラスタの粒子組成タイプで全粒子の約 70% を説明でき、上位 500 クラスタで、99% を説明できていた。本観測データの解析では、全粒子の 90% を説明できる上位 230 クラスタに注目し、残りの 10% に属する $\text{PM}_{2.5}$ 粒子は“231 番目のクラスター”のメンバーとして扱い、更にデータ解析を進めた。

3.3 クラスタリングにおける数学的な重要さと大気化学的な重要さの違い

これまでの単一微粒子質量分析計による $\text{PM}_{2.5}$ 粒子の大気観測の経験から、1 ヶ月程度の観測であれば概ね 20-50 種類の組成タイプで全粒子の 90-95% 程度を説明できる場合が殆どである。Vigilance Factor (類似性判断の閾値) を小さく値に設定し ART-2a クラスタリングを行えば、20-50 種類のクラスターに集約させることは可能である。しかし、Vigilance Factor を小さな値に設定することは、各質量スペクトルに見られる“強度が高い主要なイオンピークのパターン”だけに注目してクラスタリングを行うことであり、「イオン強度が低いが大気化学的な見地から非常に重要であるイオンピーク (化学組成) の情報」を見逃してしまうことになる。主要な粒子組成タイプを知るなどの目的には十分な解析アプローチかもしれない。しかし、これでは微粒子 1 個ごとに多様な化学物質を一括して測定でき、非常に豊富な化学情報を与えてくれる個別粒子組成測定のメリットを十分に活用できない。質量スペクトルに含まれるイオンピークの単なる強度

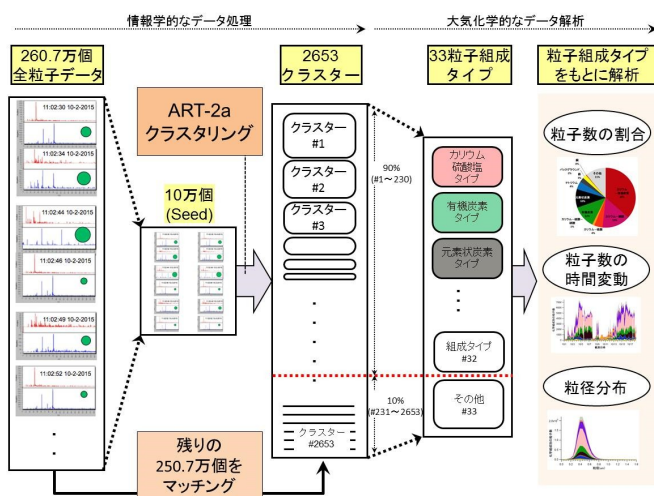
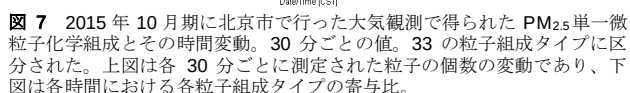


図 5 現在行っている ART-2a クラスタリングを用いた北京市での大気観測で取得した 260.7 万個の個別粒子質量スペクトルデータのデータ解析の流れ。最終的に 33 粒子組成タイプに集約された。

時間変動や粒径分布も勘案しながら、人による更なる分類によって 230 クラスターは 32 の粒子組成タイプにグループ化された。231 番目以降の全てのクラスターを1つの“その他の粒子組成タイプ”として扱い、合計 33 の粒子組成タイプに対して、大気化学的な見地からの解析を進めた。

10月3日から8日まで、北京市ではPM_{2.5}質量濃度が120-150 ng/m³に達するPM_{2.5}大気汚染に見舞われた(日本の環境基準は1年平均15 ng/m³以下かつ1日平均35 ng/m³)。PM_{2.5}濃度の増加に対応して、単一微粒子質量分析計で測定された粒子数も増加していることが分かる。まず、どのような化学組成を持った粒子組成タイプが増減しているか注目してみる。PM_{2.5}濃度が高い場合には、有機物質を多く含むタイプ(OCタイプ)と元素状炭素を多く含むタイプ(ECタイプ)が増えていることが分かる。一方、PM_{2.5}濃度が低い10月9日-11日では、バイオマス燃焼あるいは石炭燃焼の指標であるカリウム(K)を含む粒子タイプの寄与比が大きくなっていることが分かる。これらの結果から、高いPM_{2.5}濃度現象は、低濃度時とは異なるPM_{2.5}



ごく初歩的なデータ解析からでも、以上のように $\text{PM}_{2.5}$ の放出源や生成機構について様々な情報を得られることから、多様な化学成分を1つ1つの粒子ごとに連続測定するメリット、ならびに個別粒子質量スペクトルビッグデータが持つ化学情報の豊富さが分かる。更に詳細な個別粒子質量スペクトルビッグデータの解析に加えて、気象データや $\text{PM}_{2.5}$ の原料物質の1つである大気微量ガス成分の濃度変動などと比較・検討することで、更に $\text{PM}_{2.5}$ 放出源や生成機構について理解が進むと期待される。大気化学研究のフェーズである。

[Song 1999] Song, X. H., P. K. Hopke, D. P. Fergenson, and K. A. Prather: Classification of single particles analyzed by ATOFMS using an artificial neural network, ART-2A, *Anal. Chem.*, 71, pp860-865, 1999.