環境分析化学のための統計処理法

江　口　哲　史

1　序　　論

環境分析に限らず，分析により得られるデータはその測定値のみから物事が推察されるということは限られており，しばしば測定対象にかかわるメタデータと組み合わせることでその真価が示される．中でも環境分析化学においては，大気や水，土壌のような野外サンプル，室内空気やハウスダストのような屋内の試料，野生動物，魚介類，ヒトなどの生体試料など，様々な試料が分析対象となりうるため，これといった決定版となる代表的な手法が存在しているわけではないことが，解析を行う上で悩ましい点であろう．このため，多くの場合は測定したデータの特長や分布，相関などを探索的に集計・可視化していくことがデータ解析の第一歩となるだろう．このため，ここでははじめに探索的データ解析（explanatory data analysis, EDA）について紹介する．

また，試料に紐づくメタデータと測定データを組み合わせた解析を行う場合には，測定データとメタデータが含まれるデータを特定の ID に基づき結合する必要がある．特に大気，水，土壌などの環境媒体を測定対象とする場合には，試料採取場所や採取した都道府県・市区町村に関連するデータと結合して解析を行う必要があるかもしれない．しかしながら政府統計の総合窓口 e-Stat[1] などを始めとする公共データは必ずしも解析に適したフォーマットとは限らず，それらデータの整形に多大な労力を要する場合もある．このようなデータクリーニングは解析を行う前段階における重要な手順であり，解析作業以上に労力がかかる要素であるため，データの結合やクリーニングについての解説も行う．

また，いくつかのケースを想定した統計処理についても紹介する予定ではあるが，残念ながら著者も環境分析にかかわるすべてのケースを網羅できているわけではない．このため，あくまで著者の経験に基づいた例にとどまるが，紹介するデータ解析のフローが少しでも読者の参考になれば幸いである．

Statistical Methods for Environmental Analytical Chemistry.

2　データの結合・クリーニング

2･1　序

はじめにデータクリーニングに関して紹介する．序論でも述べたように，データのクリーニング・結合は，環境分析により得られた測定値を現実と結びつける上で不可欠であり，避けることができない作業であると言える．しかしながら，公共データの多くはそのフォーマットの問題から，簡単に測定値と結びつけて解析を行うことが困難である．また，測定データそのものにおいても，検出下限値や分析値の欠損の取り扱いや，文字列データの混入など，クリーニングが解析結果に影響を及ぼしうる点は同様である．この項では主に公共データ・測定データのクリーニングにかかわる話題を提供する．

2･2　公共データの取り扱い・結合

大気や水，土壌のような野外で採取されたサンプルを解析し，その特徴を考察する場合には，採取地点にかかわるメタデータは非常に有用な材料になりうる．このような公共データを収集する際には e-Stat[1] や国土数値情報ダウンロードサイト[2]，気象庁の各種データ・資料[3] などが有用である．

しかしながら，これらのデータを単純に測定データと結合し，解析に用いるには工夫が必要になるケースがある．この原因として，これらのデータの中には，様々な目的でデータが再利用されることを想定しているにもかかわらず，紙に印刷する場合の体裁を優先したレイアウトで登録されているものがあるためである．このようなファイルはデータが実際に入っている箇所がコンピュータの想定と異なっていることや，セル結合などの影響で統計処理のためのソフトウェアにデータを読み込む時点で問題が発生しうる．例えば本来数値が入力されているべきセルに単位が同時に入力されているために，数値データが文字列になってしまい，解析ができなくなってしまうという事例などが報告されている[4]．

このように，コンピュータを使った自動集計，可視化の際にはこれらを修正する必要があるため注意が必要である．具体的には結合セルの削除や，データが入力され

ていないヘッダ行の削除や，列内の数値・文字列の確認などが特に重要な事項になるだろう．これらの作業を経た後，測定データに紐づく試料採取地点情報や採取日時などの情報に基づきメタデータを結合し，データの解析に用いることができるだろう．特に国土交通省の地図データは，試料採取地点の緯度経度情報と組み合わせることで，可視化・空間統計などの手法を行う上で有用な公共データソースとなることが期待される．

2･3　測定データの取り扱い・クリーニング

環境分析により得られた測定値についても，文字列の混入を避けること，セル結合を避けることなど，公共データ同様にこれらの点については注意すべきである．一方で，測定データに目を向けると，特有の概念として定量下限値がデータを取り扱う上で問題となりうる．これらの値は測定を行ったものの定量値としての値が得られなかったものであるため，未測定とは全く異なり，一定値未満であるという情報を持っている．このような定量下限値を含むデータに対して，統計処理を行う場合には定量下限値未満のデータは 0 や定量下限値の 1/2 など，一定の値を便宜的に代入することが行われているが，これはデータの分布を歪めることにつながるため，統計学の視点から見ると必ずしも推奨されない処理である．統計学の視点においては，定量下限未満の測定結果を含むデータは本来正規分布であるデータが定量下限の値を境になくなってしまう，左打ち切りの分布と見ることができる．そのため，測定値そのものを目的変数として分析を行う場合には tobit 回帰を用いることや，統計解析を行うためのプログラミング言語である R[5] パッケージの一つである打ち切りデータを目的変数として取り扱うことができる手法をまとめた NADA2 パッケージ[6)7)] を用いることで，分布の歪みを補正して推定を行うことができる．NADA2 パッケージに関連する解析手法の詳細については成書に詳しい[7]．一方，説明変数としてこれらのデータを用いる場合には，欠損値代入のアルゴリズムや多重代入法などを用いて定量下限値未満の値に対して欠測値補完を行うことができる．

著者は R をもちいてこれらの処理を行う場合が多い．具体的には欠測データ処理を行うための mice パッケージ[8]に，qgcomp パッケージ[9]に含まれている左打ち切りデータの補完を行うための関数を組み合わせ，欠測値補完を行っている．その他にもこのような欠測値補完を行うためのパッケージとして GSimp[10] など複数のアルゴリズムが提案されており，プログラムの高速化や欠損値補完の正確さ等の指標改善のため，現在も開発が進められている．

3　E　D　A

3･1　序

他の分野における分析でも同様だが，環境分析も単一の測定値だけを使って解析を行う訳では無い．これまでに解説したような公共データとの組み合わせに基づいて解析を行う場合があるだけでなく，測定対象物質が複数種あり，それらの関係を解析してみるということもあるだろう．一方，モニタリング調査などにおいては最初から研究の仮説が決まっているわけではなく，探索的に得

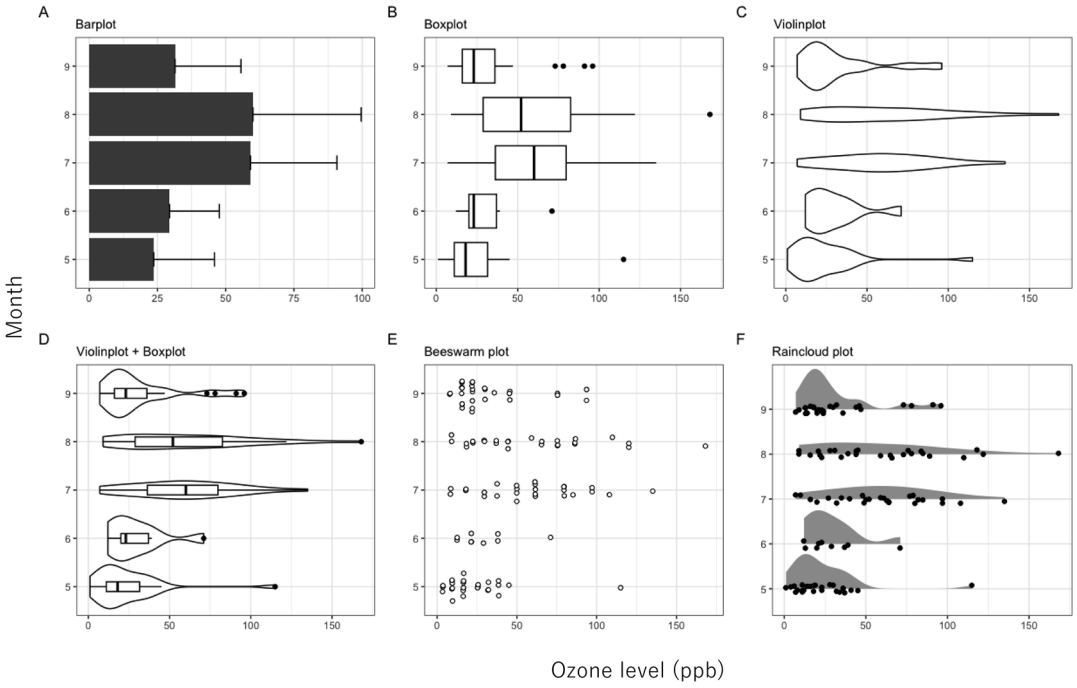


図 1　可視化法の比較，R パッケージ datasets の airquality データに含まれるオゾン濃度を月別に比較したもの，A：棒グラフ，B：箱ひげ図，C：バイオリンプロット，D：バイオリンプロット＋箱ひげ図，E：蜂群図，F：Raincloud plot

られたデータの分布などを，データの集計や可視化を通じて特長を探索することもあるだろう．このような解析はEDAと呼ばれる．

具体例をあげると測定した複数の対象物質濃度を比較する場合や，複数の地域からサンプリングした試料中の対象物質濃度を比較・可視化する際には，barplot・box-, -whisker plot・violin plot・beeswarm plot・raincloud plotなどが利用しやすいだろう．これらはいずれも濃度比較を行う際に有用な可視化手法であるが，violin plot・beeswarm plot・raincloud plotは特にデータの分布をより細かく可視化できる手法となる（図1）．また，これらの可視化を行ったデータセットについては，群ごとに測定値を比較することが次のステップになるだろう．比較のための手法を選択する際にまず考慮するべき点として，比較したい群が2群であるのか，それ以上であるのかが挙げられる．また，データの分布も同様に手法を選択する上で重要であり，データの分布が正規分布であるかそうでないか，分散が等分散であるかそうでないかが手法選択のポイントとなる．

3･2　2群の比較

初めに汚染地域と対照地域の比較や二つの測定対象物質濃度の比較を行うような，2群の比較を行うことを考える．前述の通り，群分けが定まった後はデータの分布・分散について考慮する必要がある（表1）．表1に示したように，測定したデータの分布・分散ごとに正確な検定を行うことが可能な手法は異なっている．特に，正規分布を仮定した際にしばしば利用されるStudentのt検定や，正規分布を仮定しない場合にしばしば利用されるMann-WhitneyのU検定は，比較したい変数の分布の形状が似ている（等分散）ことを仮定した手法であるため，2群のデータの分布の形状が似ていない不等分散である場合には適切な手法とは言えない．このように，2群のデータの分布の形状がことなるために不等分散が仮定される場合には，Welchのt検定やBrunner-Munzel検定[11]の利用を検討されるとよいだろう．Rにおいては Welchのt検定はt.test関数において不等分散であることをvar.equal=Fという引数で示すことで，Brunner-Munzel検定はlawstatパッケージ[12]やbrunnermunzelパッケージ[13]を利用することで実行可能である．また，複数の地点において異なる時間に2回サンプリングを行い，初回と2回目のサンプリングにおける2時点の差を検定したい場合には，同一地点で採取したペアの測定値が存在することになる．このような場合には対応のあるt検定（正規分布を仮定）やWilcoxonの符号順位検定（正規分布を仮定しない）を利用する必要があるため注意されたい．最後に，解析結果を見てから仮定（データの正規性，等分散性など）を変更することは研究倫理上推奨されないため注意すべきである．

表1　正規性・分散に基づく検定の使い分け

	正規性あり	正規性なし
等分散	Studentのt検定	Mann-WhitneyのU検定
不等分散	Welchのt検定	Brunner-Munzel検定

3･3　多群の比較

多群比較で思い浮かべる手法というと，おそらく分散分析（正規分布を仮定）やKruskal-Wallis検定（正規分布を仮定しない）を想像される方が多いのではないだろうか．これらの手法はもちろん3群以上の比較を行う際に用いる手法ではあるのだが，これらは複数群の群組み合わせのいずれかに差があるのかを検定するものであり，比較したい群のペアそれぞれの間に差があるのかを検定するわけではない点には注意が必要である．3群以上のデータを各ペアごと検定する際に，例えばt検定などを繰り返すことは推奨されない．これは有意水準を0.05にしても，検定を複数回行うと帰無仮説が棄却される可能性が高くなってしまうためである（ABCの3群の場合：AB，AC，BCの組み合わせで3回検定を行うと0.14）．そのため，p値の補正を行うか，多群の比較に用いる多重比較の手法を利用する必要がある．

多重比較の手法を使って3群以上の各ペアそれぞれに差があるのかを検定したい場合にも分布・分散の仮定が存在する点は2群比較の際と同様である．正規分布・等分散を仮定する場合にはTukey-Kramer法，正規分布・不等分散を仮定する場合にはGames-Howellの方法，正規分布を仮定しない場合にはSteel-Dwass法がしばしば利用される．これらの手法についてもいずれもRパッケージを利用することで実行可能である．

3･4　相関分析・関係解析

特定の対象物質とメタデータ・あるいは同時に測定した複数種の対象物質濃度の関係を解析する際に行うのが相関分析であろう．相関分析においても，正規性について考慮する必要があり，正規分布を仮定する場合にはピアソンの積率相関係数，正規性を仮定しない場合にはSpearmanの順位相関係数がしばしば利用される．これらの手法を用いて相関分析を行う場合には，データの可視化と相関分析を同時に行うことができると見通しが良い．また，二つの解析対象の関係を解析する場合であれば問題ないだろうが，同時に測定を行っている多数の化学物質濃度やメタデータとの関係を解析したい場合に，一つ一つ作図・解析を行うのは手間である．このような場合には，散布図行列や相関行列の作図を行うことで，ある程度の数までの変数であれば相関分析・作図を見通しの良い形で行うことができるだろう．著者はこれらの

相関分析や作図に，R パッケージである GGally[14] を主に使用している（図 2）．

一方，散布図行列は変数が増えすぎるとそれぞれの散布図が小さくなりすぎてしまうため，視認性が落ちてしまうという問題がある．このような場合には，主成分分析（principal component analysis, PCA）を用い，多変量な高次元データ（次元は解析対象の変数の数である）の次元を 2-3 次元の理解しやすい大きさの次元まで縮約することで互いの関連する変数を探索することが有用である．PCA を実行すると，試料に含まれる成分の情報が出力されるローディングプロット（図 3A）と，試料に関する情報が出力されるスコアプロット（図 3B）の二つの図が出力される．この二つの図は位置関係が互いに関連しており，スコアプロットの右側に配置された試料中ではローディングプロットの右側に配置された成分の寄与が大きいという関係が成り立つ．また，ローディングプロットで同じ向きに矢印が伸びている対象物質は互いに正の，逆向きに矢印が伸びている対象物質は負の関連を示し，直行する対象物質とは関連がないという関係も成り立っている．図 3 の場合，図 3B 右側中央にプロットされた試料 1（丸囲み）ではカドミウム，鉄，鉛が，図 3B 左側側上部にプロットされた試料 410（三角囲み）などではナトリウム・クロライドの濃度が高いことがわかる．さらに，全測定データにおけて，カドミウム等とナトリウム等の間の関連性は小さいことが示唆される．このような特長から，PCA を用いることで，同時に測定を行った複数種の対象物質間の関係を図示し，解析することが可能になる．PCA は R に初めから入っ

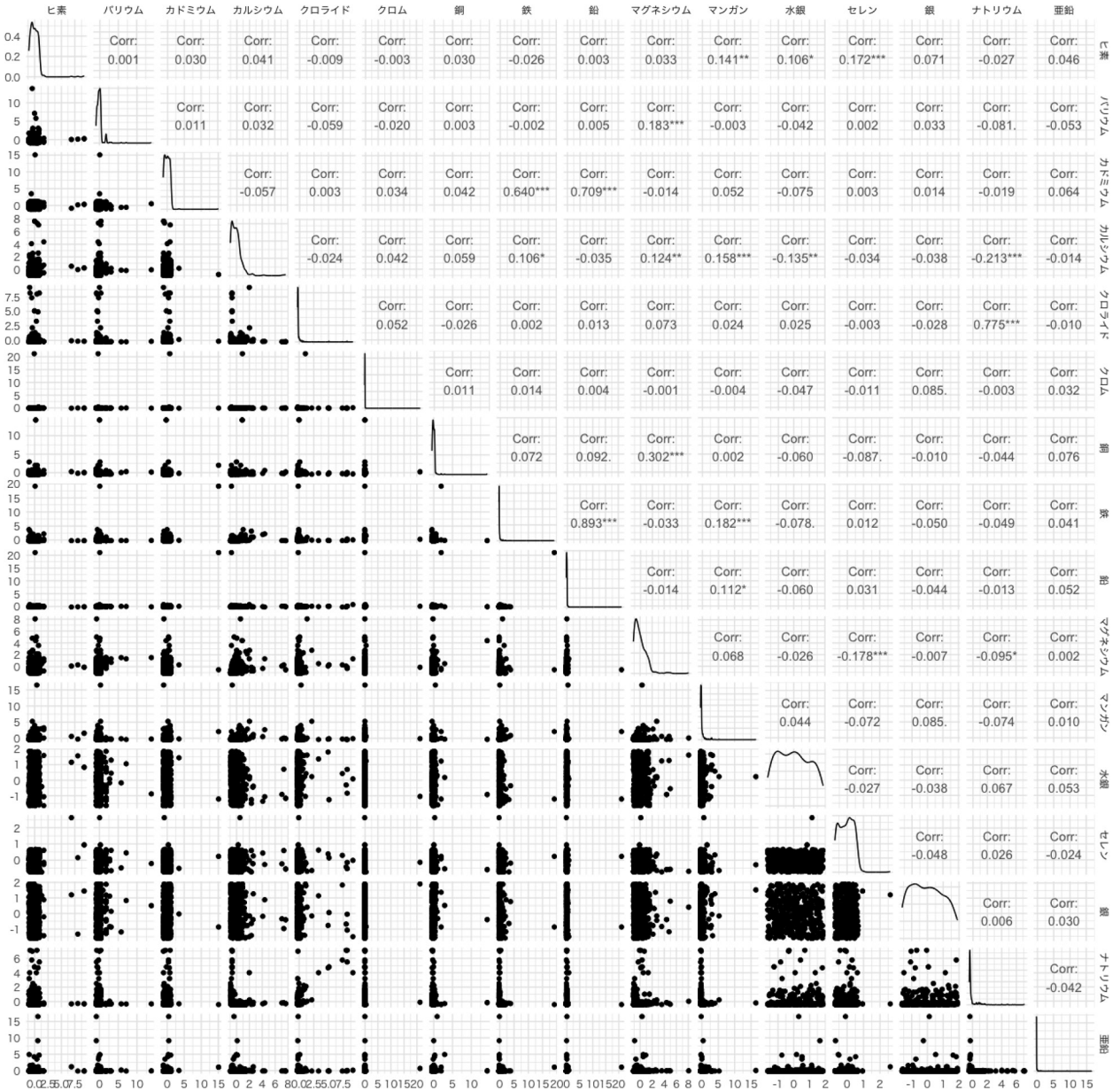


図 2　散布図行列の例

qgcomp パッケージ[6] に含まれるノースカロライナ州における井戸水中金属類濃度をシミュレートした値の散布図行列（金属類名著者訳）．すべての連続値は平均 0，標準偏差 1 に標準化済み

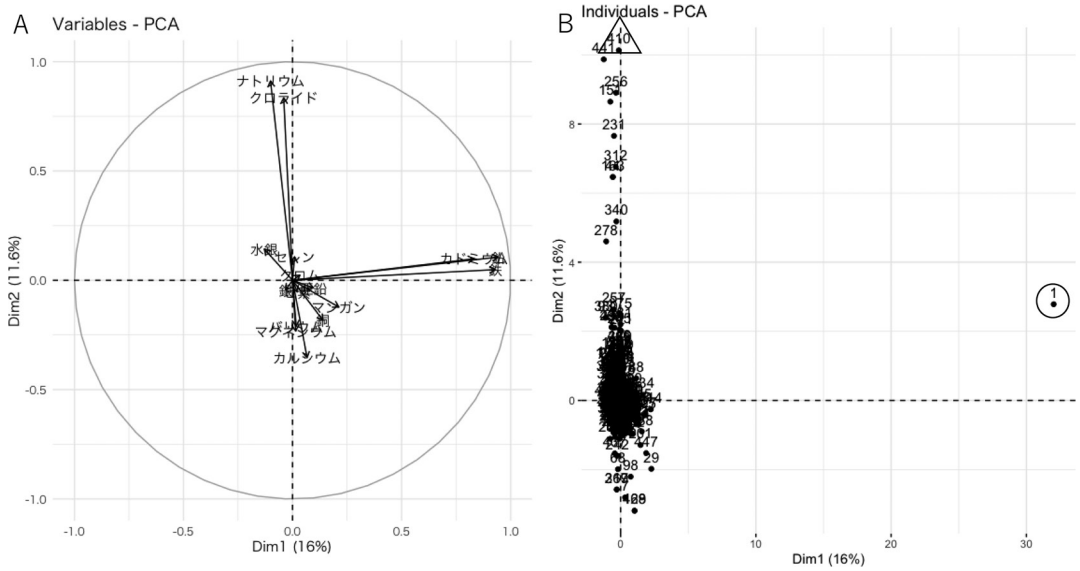


図 3　主成分分析の例

qgcomp パッケージ[6] に含まれるノースカロライナ州における井戸水中金属類濃度をシミュレートした値について主成分分析を行ったもの（金属類名著者訳）．
A：試料に含まれる成分の情報が出力されるローディングプロット；B：試料に関する情報が出力されるスコアプロット；*◯：試料 1，△：試料 410.

ている prcomp 関数を使うことで実行できるが，論文に掲載するクオリティの図を作ることを考える場合には，FactoMineR パッケージ[15] と factoextra パッケージ[16)17)] を組み合わせて利用すると良いだろう．

4　回 帰 分 析

4･1　序

環境分析により得られたデータを使って回帰分析を行う場合には，測定対象物質そのものの濃度の予測や濃度に関連する事象の影響を解析するために，測定対象物質を目的変数としてモデリングする場合と，環境・生体試料中に残留する化学物質の曝露影響を推定する場合のように，説明変数として測定対象物質を設定してモデリングする場合の 2 通りが主に考えられるだろう．前者の場合には対象物質濃度が統計的にどのような分布をしているかや時間・空間の要素が解析に含まれるかなどの研究のデザインに，後者の場合にはそれらに加え，解析に用いる対象物質が単一ではなく，多変量のデータになっていることがありうる点に注意が必要である．本節ではこれらの項目について解説する．

4･2　測定対象物質濃度を目的変数としてのモデリング

環境中の測定対象物質濃度を目的変数として回帰分析を行う場合には，対象物質濃度に関連しそうな因子から予測値を算出することや，関連する因子の探索・影響の大きさなどを推定することが主な目的になることが考えられる．対象物質濃度を目的変数とする場合，測定データには 0 未満の負の値が存在しないことが分布の特徴として挙げられる．また，汚染地域・汚染個体などをサンプリングした場合に高濃度の試料が稀に含まれることから，必ずしも分布が正規分布に従う訳では無い点に注意が必要である．この場合には，負の値を含まない分布である対数正規分布や，ガンマ分布などの分布を仮定することで，より目的変数の分布に適したモデリングを行うことができる可能性がある．

また，同一地点・同一個体から継続的にサンプリングを行うような，時間が時系列データを分析する場合や，一定の区画から複数ポイントサンプリングを行い，区画内の対象物質の空間的な分布を解析したい場合，河川水を上流から下流にかけてサンプリングするような，連続性を持つ試料のモデリングを行う場合には注意が必要である．これは一般的な線形回帰分析はそれぞれの値が独立した値であることを仮定しており，近い時間や近い地点で採取された試料同士は似た傾向を持つことを無視しているためである．このため，時系列解析や空間モデルを適用することが環境試料のモデリングを行う上で適切である場合がある．とはいえ時系列・空間モデルにも様々な手法があるため，目的に合わせた解析手法の選択を慎重に行う必要があるだろう．環境データの時系列データについては，必ずしも毎年経年でサンプリングを行うことができない場合などもあるため，データ取得の時期が不定期な場合であっても適用可能な KFAS パッケージ[18] などを用いた状態空間モデルが有用な可能性がある．また空間モデルの中でも，特定地域における化学物質の分布予測には Kriging 法を利用することで，空間相関を考慮したモデリングを実施可能である[19]．また，人口密集地とそうでない地域において汚染源が一つ増える際に及ぼす影響が異なるように，対象とした場所に

よって回帰係数が違う可能性があるような場合には，geographically weighted regression を利用することで地域性を考慮したモデリングを行える可能性がある[20]．これら時間・空間モデルについては成書に詳しい解説があるため，それらを参考にされるとよいだろう[21)22)]．

4･3　測定対象物質濃度を説明変数としてのモデリング

測定対象物質濃度を目的変数とする場合とは異なり，今回は目的変数の分布によって手法を選択していくことになるだろう．例えば身長や体重などが目的変数であれば正規分布や対数正規分布を仮定した線形回帰分析，病気の有無などであればロジスティック回帰分析が初めの候補となるだろう．測定対象物質が単一であれば，共変量を加えてモデルを組み立てることになるだろう．しかし，測定対象物質が複数種存在しており，それらすべてをモデルに組み込みたい場合には注意が必要になる場合がある．類似構造を持つ化学物質や，用途の近い化学物質を一斉分析している場合，測定されたそれらの値は互いに相関している場合がありうる．このような場合には多重共線性のために，回帰係数の推定が不安定になってしまうため注意が必要である．

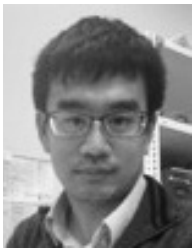
また，測定した試料の数以上に測定対象項目がある場合には，連立方程式を解く際に解が無数にある場合と同様に解が不定になってしまうため，通常の重回帰分析を行うことができない．対策として，いずれかの化合物だけを推定に用いる方法の他，部分最小二乗回帰（partial least squares regression）分析を用いることで，説明変数である測定対象物質濃度に相関がある場合にも解析を実施することができる．また，測定対象項目を混合物として曝露（ばくろ）の影響を評価したい場合のために，近年では weighted quantile sum regression[23)] や Quantile G-computation[9)] などの手法が開発されており，活用が期待されている．

5　終わりに

目的に応じて解析手法を変える必要がある点については他分野でも同様であるが，中でも環境分析は様々な目的に基づいて実施されるため，その選定が難しい面がある．この点については筆者も例外ではなく，データのクリーニングから手法の選択はいつも手探りからのスタートと言える．分析法の最適化同様，データの解析はゴールのない課題だが，本稿がなにかのヒントになれば幸いである．

文　献

1）政府統計の総合窓口 e-Stat : available from〈https://www.e-stat.go.jp/〉,（accessed 2023. 10. 27）.
2）国土数値情報ダウンロードサイト：available from〈https://nlftp.mlit.go.jp/〉,（accessed 2023. 10. 27）.
3）気象庁各種データ・資：料 available from <https://www.jma.go.jp/jma/menu/menureport.html>,（accessed 2023. 10. 27）.
4）奥村晴彦：情報教育シンポジウム 2013 論文集，**2013**, 93.
5）R. Ihaka, R. Gentleman : *J. Comput. Graph. Stat.*, **5**, 299（1996）.
6）NADA2 package : available from〈https://cran.r-project.org/web/packages/NADA2/index.html/〉,（accessed 2023. 10. 27）.
7）D. R. Helsel :“*Statistics for Censored Environmental Data Using Minitab R*”,（2011）,（John Wiley & Son）.
8）S. van Buuren , K. Groothuis-Oudshoorn : *J. Stat. Softw.*, **45**, 1（2011）.
9）A. P. Keil, J. P. Buckley, K. M. O'Brien, K. K. Ferguson, S. Zhao, A. J. White : *Environ. Health Perspect.*, **128**, 47004（2020）.
10）R. Wei, J. Wang, E. Jia, T. Chen, Y. Ni , W. Jia : *PLoS Comput. Biol.*, **14**, e1005973（2018）.
11）E. Brunner , U. Munzel : *Biom. J.*, **42**, 17（2000）.
12）lawstat package. available from〈https://cran.r-project.org/web/packages/lawstat/index.html〉,（accessed 2023. 10. 27）.
13）brunnermunzel package. available from〈https://cran.r-project.org/web/packages/brunnermunzel/index.html〉,（accessed 2023. 10. 27）.
14）Gally package. available from〈https://cran.r-project.org/web/packages/GGally/index.html〉,（accessed 2023. 10. 27）.
15）S. Lê, J. Josse , F. Husson : *J. Stat. Softw.*, **25**, 1（2008）.
16）A. Kassambara :“*Practical Guide To Principal Component Methods in R : PCA, M(CA), FAMD, MFA, HCPC, factoextra*”（2017）,（STHDA）.
17）factoextra package. available from〈https://cran.r-project.org/web/packages/factoextra/index.html〉,（accessed 2023. 10. 27）.
18）J. Helske : *J. Stat. Softw.*, **78**, 1（2017）.
19）E. J. Pebesma : *Comput. Geosci.*, **30**, 683（2004）.
20）I. Gollini, B. Lu, M. Charlton, C. Brunsdon , P. Harris : *J. Stat. Softw.*, **63**, 1（2015）.
21）馬場真哉：“時系列分析と状態空間モデルの基礎＝Foundations of Time Series Analysis, State Space Models : R と Stan で学ぶ理論と実装”,（2018）,（プレアデス出版）.
22）村上大輔：“R ではじめる地理空間データの統計解析入門”,（2022）,（講談社）.
23）C. Carrico, C. Gennings, D. C. Wheeler, P. Factor-Litvak : *J. Agric. Biol. Environ. Stat.*, **20**, 100（2015）.



江口 哲史（Eguchi Akifumi）
千葉大学予防医学センター（〒263-8522 千葉県千葉市稲毛区弥生町 1-33）．愛媛大学大学院理工学研究科博士後期課程修了．博士（理学）．《現在の研究テーマ》環境化学物質・メタボローム分析による化学物質曝露影響の解析．《主な著書》“実践 Data Science シリーズ データ分析のためのデータ可視化入門”（翻訳）,（講談社）．《趣味》R プログラミング，音楽鑑賞．