特集・環境のビッグデータとそのデータ解析技術

生態学におけるビッグデータ解析の手法
—生物群集解析の基礎理論—

門 脇 浩 明 *

キーワード：生物多様性，群集組成，多変量解析，希釈法，PERMANOVA

1．はじめに

環境評価を行う上で，生物群集や生態系の解析は不可欠である．ビッグデータ時代に突入した生態学では，とくに微生物群集や環境 DNA のデータ解析の重要性が増している．本稿では，それらの生物群集データ解析において基本となる①種数を比較する方法と②種組成を比較する方法の基礎理論を紹介する．種数などの生物多様性の指標は「スケール」に依存する値であるため，調査努力や観測スケールを考慮した上で評価することが基本である．また，群集組成などの生物多様性の指標は「多次元」であるため，その性質を適切に考慮したうえで統計的な枠組みを用いなければならない．生態学におけるビッグデータ解析における今後の挑戦と可能性について展望する．

2．生物群集解析の目的

生態学では，遺伝子配列から，生物多様性（種の組成や個体数），生態系の物質量やその流れまで，様々な種類のデータを収集してきた[1)]．例えば，生物多様性のデータ解析では，生物群集の構造（種数や種組成など）を調べたり，その構造に影響するプロセス（環境要因や生物種間の相互作用など）を明らかにしたりする．生態系のデータ解析においても，リモートセンシングやセンサーを用いて測定した生態系データから生物の活動や機能，物質の流れを推定したり，予測したりする．飛躍的な技術の発展により，様々な時空間スケールの大規模データが得られるようになったことは，生態学においても課題解決への道が開かれたと同時に，各研究者レベルでビッグデータを解析する技術が求められる時代になりつつあることを示している．

本稿では，生物群集のデータ解析，特に一般的に用いられている生物群集のメタバーコーディング・データの統計解析手法について紹介する．メタバーコーディングとは，ハイスループット DNA シーケンサーを用いることで，様々な生物に由来する DNA の配列を同時に取得し，同定する技術のことであり，微生物群集や環境 DNA の調査に活用されている[1)]．例えば，水をすくってその中に浮遊する DNA を調べることで，どのような種類の魚が生息しているのかを解明できるのが環境 DNA 分析である[2)]．

生物群集のデータは，①行をサンプル識別番号，列を種の識別番号とするメインデータ，②サンプルごとの環境要因や採集日時，場所などのデータの 2 つのデータセットからなる．メインデータの個々のセルには，どの種（あるいは operational taxonomic units; 操作的分類群単位，略して OTU）がどのサンプルにどれだけの量（生物のバイオマスや個体数，あるいは，DNA シーケンスのリード数など）検出されたのかが記録されている．メタバーコーディングを用いる場合，シーケンサーが出力した塩基配列のデータなどの一式を統計解析に使用可能なデータ形式に変換するまでのパイプラインが開発されている．こうしたバイオインフォマティクスの手法は，常に内容自体が更新されるため，最新の情報を各自確認することを求めると同時に，ここでは扱うことはしない．

一般に，種数などの生物多様性の指標は，調査

Big Data Analysis in Ecology: Basic Theory of Ecological Community Analysis
京都大学学際融合教育研究推進センター森里海連環学教育研究ユニット（京都大学フィールド科学教育研究センター）
Kohmei KADOWAKI

努力や観測スケールを考慮した上で評価しなければならない[3]．また，群集組成などの生物多様性の指標は多次元であるため，その性質を適切に考慮した統計的な枠組みで解析する必要がある[3]．これらの観点から，生物群集データの構造や特徴を説明し，希釈法（rarefaction）を用いた多様性の比較，および，PERMANOVA（permutational multivariate analysis of variance）を用いた群集組成の比較法について解説する．**図1**は生物群集データ解析のフローチャートであり，以下ではこのチャートの流れに沿って解説する．これらの解析を行うことで，生物群集の構造を決定づける環境要因や相互作用などのプロセスを推定することができる．

3．多様性を比較する（希釈法）

サンプルごとの多様性（生物の種数やOTU数）を比較することは，データ解析の第一歩である．ここで注意すべきは，サイトごとに調査努力（観察時間数や採集した個体数）が異なる場合，サンプルの多様性をそのまま比較できない点である[3]．なぜなら，より多くの個体数を観察すれば，その中により多くの種が含まれるのは当然だからである．DNAメタバーコーディングでも同様の現象が生じうる．メタバーコーディングのデータでは，サンプルごとに各種（各OTU）のリード数を集計した表として得られる（**表1**）．それは，

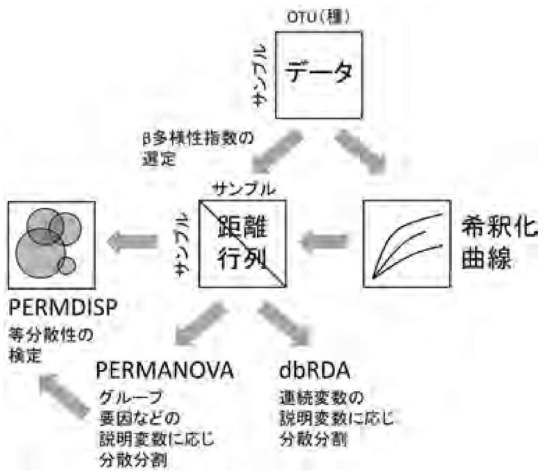


図1　データ解析フローチャート
多様性（種数）を比較する場合は，サンプル×OTU行列を用いて希釈化を行う．種組成を比較する場合は，β多様性指数を用いて距離行列に変換し，多変量解析を行う．

表1　生物群集サンプルの模擬データ

サンプル	種1	種2	種3	種4	種5	種6	種7	種8
A	35	30	30	30	1	1	1	1
B	50	40	40	10	10	5	0	0
C	35	35	30	30	24	1	0	0

シーケンサーにより，どのサンプルにどれだけ多様な生物DNAの配列数（リード数）が検出されたのかを示すものである．リード数はシーケンサーの探索努力に似ており，リード数が多いサンプルほど多くの種を含む可能性が高まるため，サンプル間でOTU数は単純に比較できない．

サンプルごとの調査努力を揃えるためにはどうすべきだろうか．それをわかりやすく説明するために，カラフルな色のついた粒状のチョコレートを例に挙げたい．1粒を1個体の生物とし，色ごとに異なる種を表すとする．そこで，サンプルの種数（OTU数）を推定することは，n個選んだ場合，そこに合計何色（何種）が含まれているのかを求めることに等しい．S種からなる群集を考え，種iの個体数N_iとし，総個体数をNとすると，サンプルサイズnのもとで期待される種数を求める．群集からn個体を選択した場合，そのn個体のうちに含まれると期待される種数Jは，ある種が少なくとも一個体は選択される確率を計算し，群集を構成する全ての種について，その確率を足し合わせた値として定義される[4]．

$$J=\mathrm{E}(S)=\sum_{i=1}^{S}\left(1-\left[\frac{\binom{N-N_i}{n}}{\binom{N}{n}}\right]\right)$$

上式を解析的に扱いやすくするため，［ ］内の組み合わせ関数を以下のガンマ関数の公式を用いて変換する．

$$\Gamma(z)=\int_0^{\infty} t^{z-1}e^{-t}dt$$

$$\Gamma(z+1)=z!$$

そのうえで，上式の両辺対数をとると，次式を得る．

$$\log(J)=\log\sum\left\{1-\frac{\Gamma(N-N_i+1)\Gamma(N-n+1)}{\Gamma(N-N_i-n+1)\Gamma(N+1)}\right\} \quad (1)$$

両辺を n について微分すると，対数の性質を利用して，以下のように変形し，種数の期待値の変化率を求めることができる．

$$\frac{1}{J}\frac{dJ}{dn}=\frac{d}{dn}\{\log(\Gamma(N-n+1)-\log(\Gamma(N-N_i-n+1))\} \quad (2)$$

$$\frac{dJ}{dn}=\left\{\frac{\Gamma'(N-n+1)}{\Gamma(N-n+1)}-\frac{\Gamma'(N-N_i-n+1)}{\Gamma(N-N_i-n+1)}\right\}\mathrm{J} \quad (3)$$

式(1)を変形して式(3)へ代入すると，種数の期待値曲線(これを個体ベースの希釈化曲線と呼ぶ)の接線の傾きを求める式が得られる．

$$\frac{dJ}{dn}=\left\{\frac{\Gamma'(N-n+1)}{\Gamma(N-n+1)}-\frac{\Gamma'(N-N_i-n+1)}{\Gamma(N-N_i-n+1)}\right\}e^{\{\log(\Gamma(N-N_i+1)+\log(\Gamma(N-n+1))-\log(\Gamma(N-N_i-n+1))-\log(\Gamma(N+1))\}}$$

あるリード数における接線の傾きは，そのリード数における多様性の探索努力の指標（もしくは完遂度）と考えられる．種数をサンプル間で比較する最も公平な方法は，全てのサンプルについて，等しいリード数ではなく，それぞれのサンプルで等しい傾きが得られるリード数（カバレッジ）において，多様性を比較する方法である[5]．種数の増加が頭打ちになったサンプルのみを解析に用いることもできるが，それは十分なリード数が得られなかったサンプルを統計解析から除外することになり，無駄が生じる．**図2**(a)は，**表1**の模擬データを用いて希釈化を行った結果を示している．サンプル間の調査努力の違いによって，種数の推定結果が逆転する．だから，調査努力を適切に考慮することが重要なのである（**図2**(b)）．

4．群集組成を比較する

群集組成を比較するには，群集構造の非類似度を定量化する必要がある．データからサンプル間の群集組成の違いを解析するための鍵となるのが，距離行列（distance matrix）である．距離行列とは，サンプル間で定義される距離を配列して，行列として表示したものであり，サンプル数がN個あれば，その距離行列はN行N列の対称行列となる．Gower（1966）[6]は，どのような距離行列であっても，ユークリッド座標系において線形の形で表現できることを示した．すなわち，多次元のサンプル・データを点としてユークリッド空間上に表現し，それら点間の類似度の大きさの大小関係をその空間上の点間距離（位置関係）で表すことができるのである．この手法は，主座標分析（principal coordinate analysis; PCoA）と呼ばれ，多次元尺度構成法の基礎とも言える手法である[7]．2000年以降，Gower[6]の知見が基礎となり，統計学者らのさらなる尽力により，回帰分析や分散分析などの従来の統計学の整合性を有する形で，多次元データ解析の枠組みが完成された．

その突破口となったのは，McArdleとAnderson

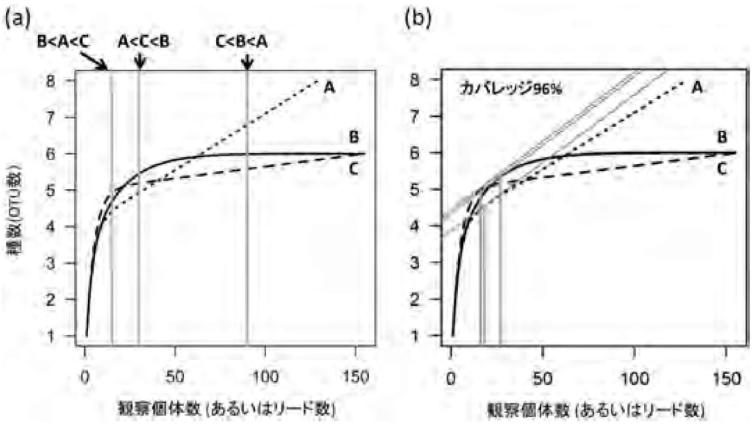


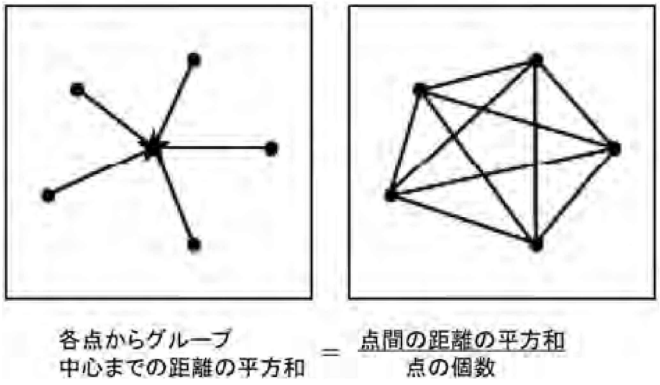
図2　希釈化曲線

(a)どれだけの個体数を観察するのかによって，**表1**の3つのサンプルのうち，いずれのサンプルの種数が高いのかという推定結果が，希釈化曲線どうしの交点を機に逆転する．(b)希釈化曲線に傾きが同じ値になる箇所で接線を引き，その接点のリード数を基準とすることで調査努力の調整を行う．カバレッジ96%（傾きが0.04）のリード数において種数の比較を行った結果，サンプルBが最も多様性が高いという結論になる．Chase et al.(2018)を参考に，Rのパッケージveganのrareslopeという関数を用いて作成．

(2001)[8]であった．彼らは，距離行列の誤差二乗和と十字積行列を関心のある説明変数(正確には，計画行列）に応じて分割できることを数学的に証明した．その証明を可能にしたのが，ホイヘンスの定理である[9]（**図３**）．すなわち，あるグループのサンプル間の平均距離は，距離の二乗平方和をすべてのサンプルの組み合わせについて足し合わせ，それをグループ内のサンプル数で割った値に等しい．この定理に従うと，分散分析と同様，中心（もしくは重心）がわからなくても距離行列だけで，計画行列に対する分散の分割が可能となる（**図４**）．その証明には，正規方程式の導出[10]やGower の中心行列[6]を理解することが必要となる．

さらに，McArdle と Anderson[8]は，説明変数の効果を並べ替え検定によって検定できることを示した．サンプルに割り当てられたグループのラベル（処理区の識別記号）をシャッフルして擬似Ｆ統計量を繰り返し計算し，実測値と比較することで，グループ（処理）の影響が統計的に有意であるのかを検定できる．距離指数としてユークリッド距離を用いる場合，この統計量はフィッシャーのＦ統計量と同値である．

この多次元データの解析手法を，広く利用可能な形に確立したのが PERMANOVA である[9, 11]．これは特定の統計分布を仮定しない手法であるため，汎用性が高く，総計ソフトＲでも利用でき，生態学にとどまらず様々な科学分野で広く使われる多次元データ解析の手法となっている．実際の生物群集の解析に対する PERMANOVA のいくつかの適用例は，門脇 (2016)[12]で解説されている．



（Anderson（2001）をもとに作成）

図３　ホイヘンスの定理の概念図
主座標分析のプロット上にデータ点(黒丸)と重心(☆印)を示している．

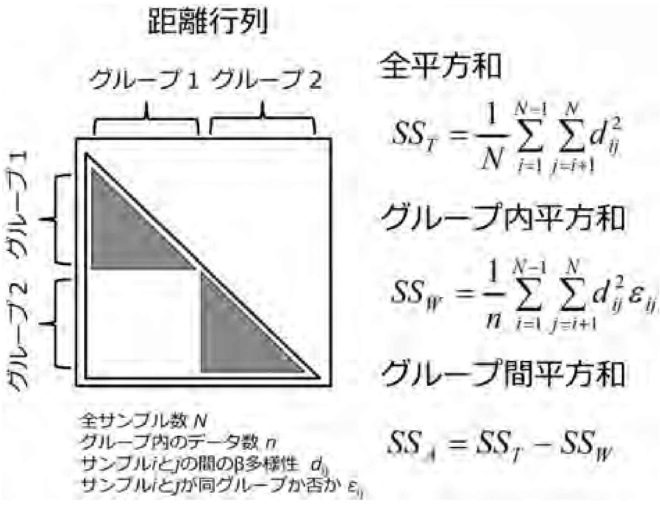


図４　二要因分散分析を PERMANOVA で行う場合の解析方法

５．PERMANOVA を用いた解析

PERMANOVA は，処理区や環境要因ごとに群集構造が異なるかどうかを検定できるパワフルなツールであり，ランダム効果や層別化（入れ子状のデザイン）など複雑な実験デザインにも対応可能である[11]．ただし，その利用にあたっては，いくつかの注意点が必要である．

第一に，距離行列を作成する際，非類似度指数を用いるが，絶対的かつ普遍的に正しい距離の測り方は存在しない．非類似度指数は，生態学では β 多様性指数とも呼ばれ，その種類は非常に多い[7, 13]．数ある指数の中からどの指数を用いるべきかについては，基本的に「多様性のどのような要素を比較したいのか」ということに尽きる（より詳細なガイドラインは Anderson et al.（2011）[14]を参照）．β 多様性指数の選択は PERMANOVA の結果に大きな影響を与えるため，目的に合わせ，適切に行う必要がある．

第二に，PERMANOVA は多次元データの相関に非常に強い一方で，データのばらつきの異質性には弱い．よって，比較する処理区ごとにサンプル数が異なる不公平な実験デザインには適さない．そのため，等分散性の検定（PERMDISP）を行うか，等分散

性を仮定しない改訂版のPERMANOVAの検定を行う必要がある[11]．なぜなら，平均的な群集構造が異ならない場合でも，グループ間で分散が異なれば，PERMANOVAは有意なグループ間の違いを検出してしまうからである（**図5**）．PERMDISPは，サンプルが相対的にどれほど広がっているのかを推定するものであり，その広がりの方向性（データ点塊の空間的な形状）は推定しているわけではない[9,11]．PERMANOVAは，高次元データに対しても有効であると考えられているが，その有効性は十分に検討されておらず，多変量誤差を推定する手法を用いて吟味するべきである[15]．

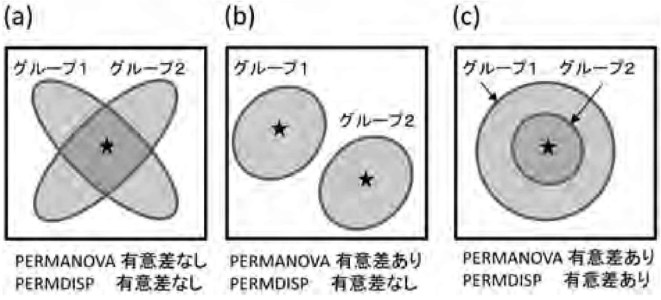


図5 2つのグループの群集構造のパターンと，PERMANOVAとPERMDISPの結果の対応関係
主座標分析のプロット上において2つのグループの群集構造を比較する場合を示している．

6．今後の展開

近年では，オミクス，画像や音声，センサーのデータなど，生物群集や生態系に関する様々なタイプのビッグデータが蓄積されつつある．このデータを解析することで，生物群集の構造と動態を支配するプロセスを理解したり，生物群集の環境変化に対する応答を予測したりすることができるだろう．また，リアルタイムデータを入手することが容易になっており，時系列データのモデリング，なかでも，フィードバックの解析をさらに深化させていくことができるかもしれない．生態学では，近年，観察や実験により，遺伝子から生態系まで様々なスケールで生命現象のフィードバックが生じ得ることが明らかになっており，フィードバックの構造や方向性を時系列データから抽出することは今後の重要な課題となっている．

謝辞：本研究は，科研費（13J02732，17K15284），ならびに日本財団・京都大学共同事業「森里海連環再生プログラム」の支援を受けた．

参 考 文 献

1）福森香代子，門脇浩明；生態系・生物群集解析法（テクニカルノート2），『遺伝子・多様性・循環の科学：生態学の領域融合へ』（門脇浩明，立木佑弥編）京都大学学術出版会，2019.
2）山中裕樹，源利文，高原輝彦，内井喜美子，土居秀幸；環境DNA分析の野外調査への展開，日本生態学会誌，**66**，(3)，p.601−611，2016.
3）Jonathan M Chase et al.; Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities, Ecology Letters, **21**, p.1737−1751, 2018.
4）Stephen H Hurlbert; The non-concept of species diversity: a critique and alternative parameters, Ecology, **52**, p.577−589, 1971.
5）Anne Chao A, Lou Jost; Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size, Ecology, **93**, p.2533−2547, 2012.
6）John S Gower; Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika, **53**, p.325−338, 1966.
7）小林四郎；生物群集の多変量解析，蒼樹書房，194p.，1995.
8）Brian H McArdle, Marti J Anderson; Fitting multivariate models to community: a comment on distance-based redundancy analysis, Ecology, **82**, p.290−297, 2001.
9）Marti J Anderson; A new method for non-parametric multivariate analysis of variance, Austral Ecology, **26**, (1), p.32−46, 2001.
10）ジョンソンRA，ウィッチャンDW；多変量解析の徹底研究，現代数学社，1992.
11）Marti J Anderson；“Permutational Multivariate Analysis of Variance（PERMANOVA）”，Wiley StatRef: Statistics Reference Online, https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat07841, 2017,（accessed 2019−04−04）.
12）門脇浩明；パッチ状環境における生物多様性の維持機構，日本生態学会誌，**66**，(1)，p.1−23，2016.
13）土居秀幸，岡村寛；生物群集解析のための類似度とその応用：Rを使った類似度の算出，グラフ化，検定，日本生態学会誌，**61**，p.3−20，2011.
14）Marti J Anderson et al.; Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist, Ecology Letters, **14**, p.19−28, 2011.
15）Marti J Anderson, Julia Santana-Garcon; Measures of precision for dissimilarity-based multivariate analysis of ecological communities, Ecology Letters, **18**, p.66−73, 2015.