技術論文

END(Enhanced Noise Discriminative)-Lassoによる重要因子同定手法の高度化：安定性評価・予測精度に基づく絞り込み・因果順序の推定

Advancing the END (Enhanced Noise Discriminative)-Lasso Framework for Identifying Key Factors: Stability Assessment, Predictive-Accuracy–Driven Refinement, and Estimation of Causal Ordering

福 島 寿 和* 中 川 淳 一 川 野 秀 一 押 木 守
Toshikazu FUKUSHIMA Junichi NAKAGAWA Shuichi KAWANO Mamoru OSHIKI

抄 録

本報告では，これまでに開発した重要因子同定手法“Lasso＋Bootstrap法”の課題である信頼度閾値の恣意性とノイズ下の再現性低下に対し，予測精度と偽陽性抑制を両立するEND (Enhanced Noise Discriminative)-Lasso枠組みを提案した。信頼度順の逐次追加で交差検証のR^2局所最大から因子群を確定し，Model-Xノックオフで候補を偽発見割合の目標上限qで絞り込み，LiNGAM＋ロバスト回帰可視化で因果方向候補を整理する。実験室規模水処理データでR^2を0.781→0.898へ改善し，直接寄与微生物の優先順位づけを可能にした。連続値・2値目的変数に適用可能である。このように，改善された本手法は水処理分野にとどまらず，今後多様な社会課題の解決に貢献し得る統計解析手法として期待される。

Abstract

We propose the END (Enhanced Noise Discriminative)-Lasso framework that simultaneously improves predictive performance and suppresses false positives, addressing two key limitations of our previously developed important-factor identification method, the “Lasso＋Bootstrap method”: (i) arbitrariness in setting the confidence (selection frequency) threshold and (ii) reduced reproducibility under noisy conditions. Specifically, we determine the final factor set by sequentially adding variables in descending order of confidence and selecting the local maximum of cross-validated R^2. We then refine candidate variables by introducing Model-X knockoffs and controlling the selection using the target false discovery rate upper bound q, and finally organize candidate causal directions by combining LiNGAM with robust regression-based visualization. Applying the proposed framework to laboratory-scale wastewater treatment data improved the predictive accuracy from $R^2=0.781$ to $R^2=0.898$, enabling prioritization of microbial candidates that may directly contribute to treatment performance. The framework is applicable not only to continuous outcomes but also to binary response variables. Overall, the improved method is expected to contribute to solving a wide range of societal problems beyond the wastewater treatment field.

1. 緒 言

著者らはこれまでに，膨大な種類の微生物が混在する排水処理プロセスの測定データに対し，重要な水処理微生物を同定する新規統計手法の開発に取り組んできた。生物学的排水処理法の代表である活性汚泥法（activated sludge process）は都市下水や工場排水の処理に幅広く用いられており，鉄鋼業においても，コークス製造工程でコークス炉より生成される通称“安水”と呼ばれる排水の処理に活性汚泥法が適用されている。この処理プロセス内に存在する微生物群は，次世代シーケンス（next-generation sequencing, 以下NGS）解析により遺伝子情報を取得することで網羅的に解析できるようになった。しかしながら，数千～万種の微生物が混在していることがわかり，複雑多様な情報から水処理に影響する重要微生物を同定することは従来困難であった。

* 先端技術研究所 環境基盤研究部 上席主幹研究員 博士(環境学) 千葉県富津市新富20-1 〒293-8511

この課題を克服するため，我々は微生物学－数理学の学融合により，Lasso＋Bootstrap 法などの新規統計手法を開発し，重要水処理微生物を同定することに成功した[1,2]。Lasso＋Bootstrap 法はスパース推定[3] の一つである Lasso（Least absolute selection and shrinkage operator）推定と，Bootstrap 法[4] による再標本化技術を組み合わせることで，信頼度という新たな指標を導入し，信頼性の高い重要微生物の絞り込みや，サンプル数が少ない場合でも解析可能な枠組みを実現した。

Lasso＋Bootstrap 法の開発により，コークス炉排水中のフェノールやチオシアン等の汚濁物質分解に関わる重要微生物を同定することに成功し，重要微生物を対象とした水処理高度化のための研究開発が進んでいる。一方で，当該手法の課題も見えてきた。一例として，信頼度は Lasso＋Bootstrap 法の特徴であるが，その閾値設定には指導原理が無く，主観的判断が入らざるを得ない。また，ノイズの影響が大きいデータの解析や，水処理速度のような定量的な値（連続値）ではなく，○×のような 2 値データの解析が必要なケースも想定される。

これらの課題や今後想定されるデータ特性に対応するため，複数の解析手法を新たに開発し，本報告ではそれらの技術について紹介する。なお，Lasso＋Bootstrap 法をはじめとする本手法群は，排水処理に関わる重要微生物の同定を目的として，水処理関連微生物を説明変数，水処理性能を目的変数として設計してきた。しかし，本手法は水処理以外の微生物群集や，さらには微生物以外の膨大な候補因子から重要因子を抽出する必要がある分野にも適用可能である。したがって，本手法は水処理分野にとどまらず，今後多様な社会課題の解決に貢献し得る統計解析手法として期待される。

2. 本　　論

2.1 モデル予測精度を指標とした重要因子群の抽出方法

既報の Lasso＋Bootstrap 法は，重要微生物候補の抽出に有効である一方，信頼度に対する閾値設定には指導原理が乏しく，解析者の主観が入り得るという課題がある。そこで本研究では，この課題への解決策の第一段階として，信頼度ランキングに基づく逐次追加モデルを構築し，予測性能（決定係数 R^2）の局所最大により重要因子群を決定する手順[5] を提案する。

本報告では，提案手順の有効性を示すため，既報[1] で取得・整理した実験室規模の流動床担体プロセス（Laboratory-scale Moving Bed Biofilm Reactor，以下 MBBR）のデータをケーススタディとして用いた。スポンジ担体を用いて，チオシアンおよびアンモニアを主成分とする人工排水の処理を行った。運転期間中は流入速度を調整して水理学的滞留時間（HRT）を変化させ，処理負荷を変動させた。処理水質を定期的に分析し，チオシアンの 1 日当たり処理速度（mg/L/day）を算出し，以降これを目的変数 y（チオシアン酸イオン（SCN^-）除去，以下 SCN 除去）とした。説明変数 X は，担体試料から抽出した DNA の NGS 解析に基づく微生物種の相対存在量であり，サンプル数 n=23，候補因子数 p=1945 である。

Lasso＋Bootstrap 法により，説明変数である微生物種ごとに選択頻度（本研究では信頼度（Selection frequency））を算出した。得られた信頼度を降順（大きい値から順）に並べたランキング結果を図 1 に示す。

次に，このランキングに基づき，信頼度の高い微生物種から順に説明変数へ 1 種ずつ追加した線形重回帰モデルを構築し，予測精度を決定係数 R^2 により評価した。説明変数の個数と R^2 の関係を図 2 に示す。

図 2 より，説明変数の追加に伴って R^2 は増加するが，ある因子数を境に増分が飽和し，さらに因子を追加すると R^2 が低下する局所的なピークが確認できる。そこで本研究では，R^2 が局所最大（あるいは最大）となる因子数 k^* を与える上位 k^* 種の集合を重要因子群と定義する。

図 3 に，図 2 で同定した局所最大点に対応する重要因子群を用いて構築した線形回帰モデルの予測精度を示す。実測値と予測値の対応は良好であり，決定係数は R^2=0.781 であった。

図 4 に，重要因子群を用いた線形回帰モデルについて，時系列での予測結果と各因子の寄与を可視化した結果を示す。上段は目的変数（SCN 除去）の実測値とモデル予測値の時系列推移である。中段はモデルに用いた説明変数（重要微生物）のサンプル間変動を示すヒートマップであり，各微生物の相対存在割合（サンプル内で正規化した値）を表す。下段は，回帰係数と中段の説明変数値（平均からの

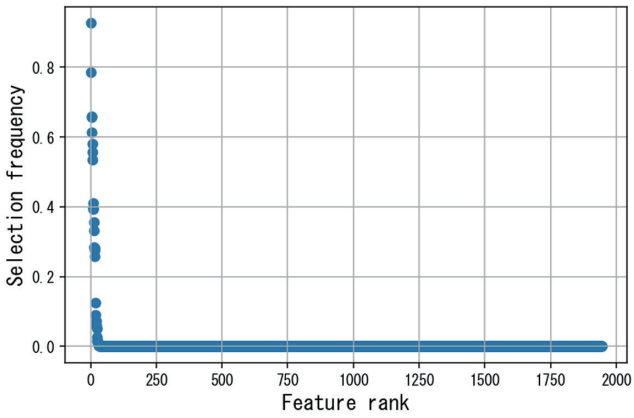

図 1　Lasso＋Bootstrap 法により算出した微生物種ごとの信頼度（選択頻度）を降順に並べたランキング結果
横軸は説明変数（微生物種）の順位，縦軸は信頼度を示す。
Ranking of microbial taxa by selection frequency computed by Lasso＋Bootstrap method (sorted in descending order) The horizontal axis shows the rank of the explanatory variables (microbial taxa), and the vertical axis shows the confidence level.

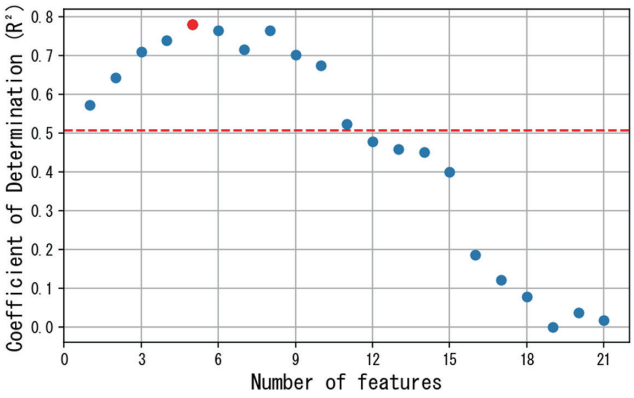

図２　図１のランキング上位から順に説明変数を追加した線形回帰モデルにおける決定係数 R^2 の推移
横軸は使用した説明変数の個数，縦軸は R^2 を示す。局所最大（赤点）を与える因子数 k^* を採用し，上位 k^* 種を重要因子群と定義する。
Changes in the coefficient of determination (R^2) of the linear regression model as explanatory variables are added sequentially from the top of the ranking in Fig. 1.
The horizontal axis shows the number of explanatory variables used, and the vertical axis shows R^2. We adopt k^*, the number of variables that gives a local maximum (red dot), and define the top k^*taxa as the set of important factors.

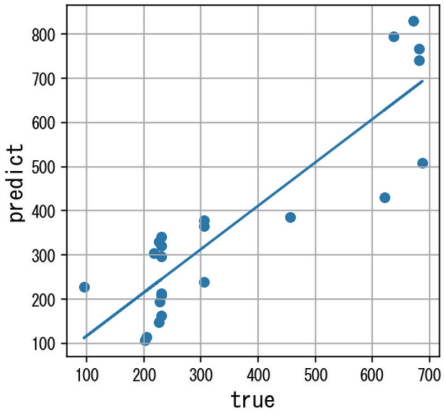

図３　図２で同定した局所最大点に対応する重要因子群を説明変数として構築した線形回帰モデルの予測結果
横軸は実測値，縦軸は交差検証による予測値である。決定係数は R^2=0.781 であった。
Prediction results of the linear regression model constructed using the important feature set (microbial group) corresponding to the local maximum identified in Fig. 2 as explanatory variables.
The horizontal axis shows the observed values, and the vertical axis shows the cross-validated predicted values. The coefficient of determination was R^2=0.781.

偏差）に基づいて算出した寄与量のヒートマップで，各サンプルにおいて予測値がどの因子によって押し上げられ／押し下げられたかを示している。これらを併せて読むことで，モデルがどの因子の変動を根拠に予測しているかを直感的に把握でき，予測精度に加えて説明可能性の観点からも重要因子群の妥当性を検討できる。

2.2　END（Enhanced Noise Discriminative）-Lasso

前節では，信頼度に対する絶対閾値を設けず，信頼度ランキングに基づく逐次追加モデルの予測性能（決定係数 R^2）の局所最大により重要因子群を決定する指針を与え，閾値設定に起因する主観性を大幅に低減した。一方で，ノイズの影響が大きいデータでは重要因子候補が多数残りやすく，因子選択の安定性・再現性が低下し得る。また，候補因子が多い状況では，ランキング上位をそのまま採用するだけでは，冗長な因子や相関の高い因子が混在し，予測精度が頭打ちになる場合がある。

そこで本節では，前節の“予測性能に基づく絞り込み”という考え方を基盤に，(i) ノイズ起因の偽陽性を抑制して候補集合を整理する工程と，(ii) 候補集合から前方逐次特徴選択（逐次追加）により予測性能が最大となる組合せを選ぶ工程を組み合わせた，重要因子抽出の一連手順（END-Lasso）を提案する。ここで用いるノイズ抑制の具体手段の一つとして，Model-X ノックオフ[6,7]に基づくノイズ変数を導入し，偽陽性を抑制しつつ，選択の安定性と予測精度の同時改善を図る[5]。

ノックオフとは，元の説明変数 X がもつ相関構造（統計的な依存関係）をできるだけ保ったまま，目的変数 y に対しては“本物の因子と同じ基準で比較できるが，因果的には不要な対照（擬似変数）”となるノイズ変数を人工的に生成し，元の説明変数と並べて学習させる手法である。このとき学習では，各説明変数と対応するノックオフ変数が同時にモデルへ入力されるため，“本物の因子”と“ノイズ因子”が同じ土俵で競争する形になる。結果として，ノイズによる見かけの相関に引きずられてモデルが過適合（オーバーフィット）することを抑制しやすくなる。

さらに重要因子の選択段階では，q（偽発見割合の目標上限）を明示的に与えることで，“採用した因子群に，どの程度まで偽陽性が混ざることを許容するか”を事前に定められる点が大きな特徴である。一般に q を小さくするほど選択基準は厳しくなり，抽出される因子数は減る一方で，偽陽性の混入が抑えられる（＝“ノイズに勝ち残った因子だけが残る”）方向に働く。一方で q を小さくし過ぎると，真に重要な因子であっても取りこぼす可能性が高まるため，本研究では“偽陽性を抑える保守性”と“予測精度”の両方を確認しながら q を設定する。

図５は，END-Lasso をブートストラップ 1000 回で繰り返し実行し，q を選択が成立する範囲で可能な限り小さく設定した条件下で抽出された 12 種の微生物候補を対象として，前方逐次特徴選択（信頼度順位に従い 1 因子ずつ追加）における“採用因子（追加順）”と“決定係数 R^2”の関係を示したものである。因子数の増加に伴い R^2 は上昇するが，ある段階で増分が飽和し，さらに因子を追加すると R^2 が低下する挙動が確認された。そこで本節では，過学習の兆候を避けつつ予測性能を最大化する観点から，R^2 が最大となる因子数で打ち切った上位因子集合を重要因子群と

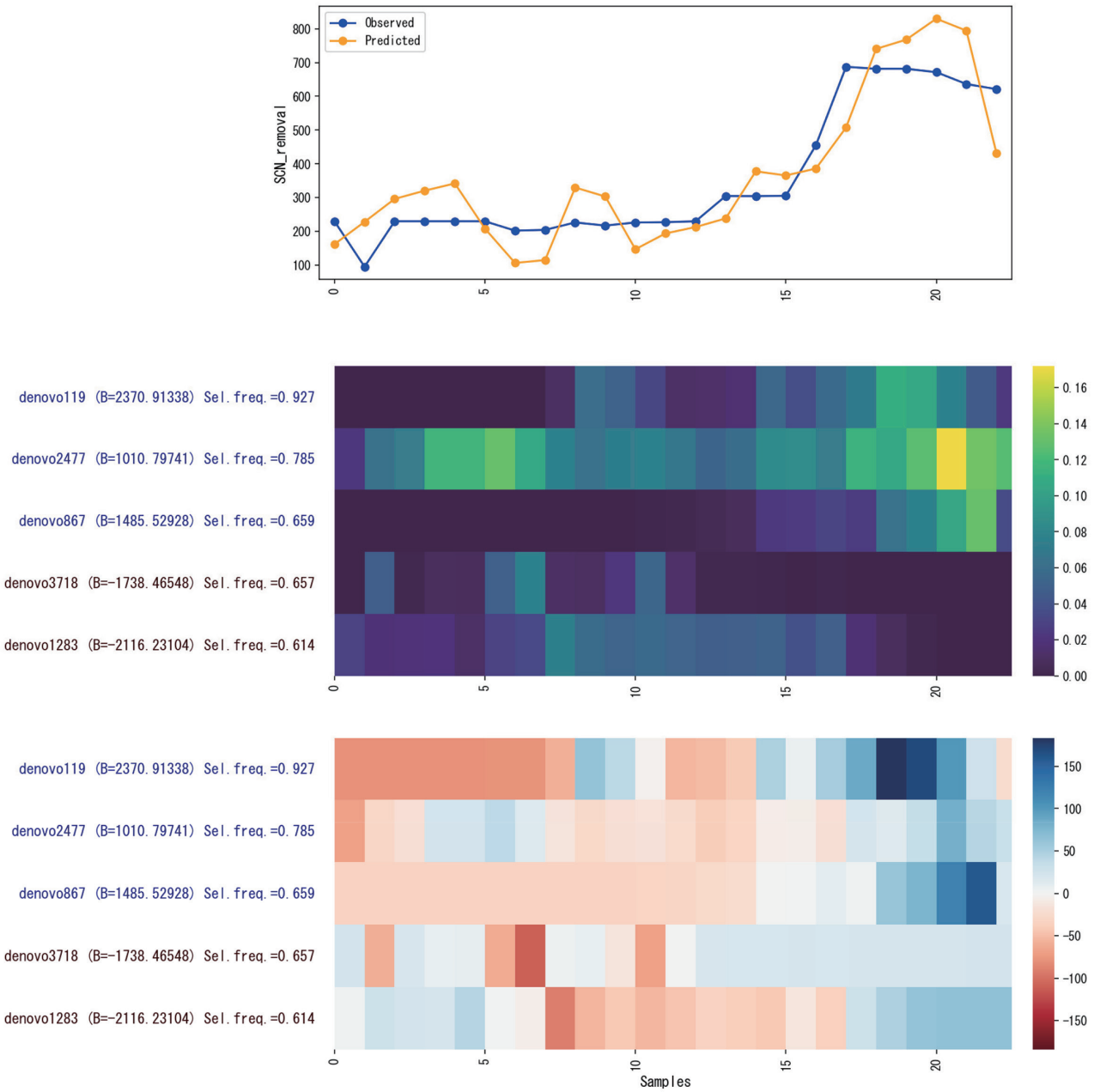

図４　重要因子群モデルの時系列予測と寄与分解の可視化

上段は，目的変数（SCN 除去）の実測値と，重要因子群を用いた線形回帰モデルによる予測値の時系列推移を示す。中段は，モデルに用いた説明変数（重要微生物）のサンプル間変動を示すヒートマップであり，各微生物の相対存在割合（サンプル内で正規化した値）を表示している。下段は，各微生物の“寄与量”を可視化したヒートマップである。寄与量は各微生物 i・サンプル s について寄与量 $=B_i\times(x_i-\bar{x})$（B_i：回帰係数，x_i：中段の相対存在割合，$\bar{x}$：その平均）で定義し，中段の相対存在割合を平均からの偏差（中心化）に変換したうえで回帰係数を掛け合わせている。したがって，相対存在割合が平均より高い（$x_i-\bar{x}>0$）とき，B_i が正なら予測値を押し上げる正寄与（青）として，B_i が負なら予測値を押し下げる負寄与（赤）として表れる（平均より低い場合は符号が反転する）。色の濃淡は寄与の絶対量の大きさを表す。なお，左側の微生物名の文字色は回帰係数の符号（正：青，負：茶）を示す。図中の Sel.freq. は Selection frequency（信頼度）の略。

Visualization of time-series predictions and contribution decomposition for the important-feature-set model.

The top panel shows the time-series trajectories of the observed values of the response variable (SCN removal) and the values predicted by the linear regression model using the important feature set. The middle panel is a heatmap showing between-sample variation in the explanatory variables (important microbial taxa) used in the model, displaying each taxon's relative abundance (values normalized within each sample). The bottom panel is a heatmap visualizing the “contribution” of each taxon. For taxon i in sample s, the contribution is defined as $\text{contribution}_{i,s}=B_i\times(x_{i,s}-\bar{x}_i)$, where B_i is the regression coefficient, $x_{i,s}$ is the relative abundance shown in the middle panel, and $\bar{x}_i$ is its mean. Thus, the relative abundances are converted to deviations from the mean (centered) and then multiplied by the corresponding regression coefficients. Consequently, when the relative abundance is higher than average ($x_{i,s}-\bar{x}_i>0$), a positive B_i appears as a positive contribution (blue) that increases the predicted value, whereas a negative B_i appears as a negative contribution (red) that decreases the predicted value (when the relative abundance is lower than average, the sign is reversed). Color intensity represents the magnitude of the absolute contribution. The font color of the microbial names on the left indicates the sign of the regression coefficient (positive: blue; negative: brown). “Sel. freq.” is an abbreviation for selection frequency (confidence).

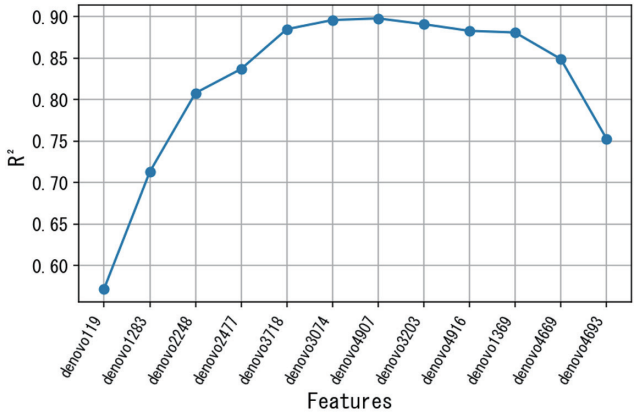

図５　q=0.085 による候補因子からの逐次追加と決定係数
q=0.085 の END-Lasso により抽出した 12 種の候補微生物因子を対象に，信頼度ランキング上位から順に 1 因子ずつ追加する前方逐次特徴選択を行い，線形回帰モデルの予測性能を決定係数 R^2 で評価した結果。R^2 は因子追加に伴い上昇するが，ある段階で最大化し，その後低下する挙動を示す。
R^2 trajectory during forward stepwise addition of candidate factors (q=0.085).
Using the 12 candidate microbial factors extracted by END-Lasso with q=0.085, we performed forward stepwise feature selection by adding one factor at a time in order from the top of the confidence (selection-frequency) ranking, and evaluated the predictive performance of the linear regression model using the coefficient of determination (R^2). The R^2 value increases as factors are added, reaches a maximum at a certain point, and then decreases thereafter.

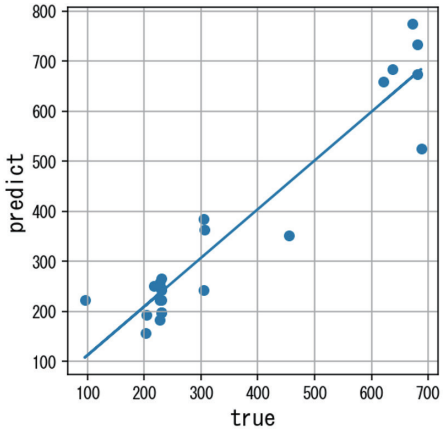

図６　重要因子群７因子による予測精度
図５で R^2 が最大となる７因子（denovo119～denovo4907）を用いて構築した線形回帰モデルの予測結果（実測値と予測値の散布図）。決定係数は R^2=0.898 を示し，前節のモデル精度 R^2=0.781 を大幅に上回る。
Predicted vs. observed values for the model built with seven selected key factors.
Prediction results (scatter plot of observed vs. predicted values) for the linear regression model constructed using the seven factors (denovo119–denovo4907) that yield the maximum R^2 in Fig. 5. The coefficient of determination was R^2=0.898, which is substantially higher than the model performance in the previous section (R^2=0.781).

定義した。その結果，12 種の候補から 7 因子（denovo119～denovo4907 の組合せ）が最適因子集合として選定された[8]。

図６に，上記 7 因子で構築した線形回帰モデルの予測結果（実測値と予測値の対応）を示す。予測は良好であり，決定係数は R^2=0.898 を示した。これは前節で得られた精度（例：R^2=0.781）を大幅に上回っており，ノックオフ変数によるノイズ峻別（q の厳格化）→ 候補因子の適切な絞り込み → 逐次追加による最適因子群の同定という流れが，予測性能の観点から有効に機能したことを示唆する。

さらに図７では，同モデルの時系列データに対する実測値と予測値の推移に加え，重要因子のサンプル間変動と，各因子の寄与（係数と説明変数値に基づく寄与量）をヒートマップとして可視化した。

2.3　重要因子群に対する因果候補の優先順位づけ

前節までに，Lasso＋Bootstrap 法により重要微生物候補を抽出し，さらに END-Lasso と逐次特徴選択を組み合わせることで，予測精度（決定係数）を最大化する観点から重要因子群をより厳密に絞り込んだ。これにより，SCN 除去の変動を高精度に再現できる説明変数集合が得られた。一方，実用上の次の課題は，“重要因子群の中で，SCN 除去に直接かかわる（原因系として働く）微生物種はどれか”を特定することである。すなわち，予測に有効な因子群の同定から一歩進めて，原因系としての候補を整理し，現場での制御・介入（運転条件の最適化，微生物叢の誘導，モニタリング指標の設計）につなげる必要がある。

ただし，観測データのみから因果関係を厳密に同定することは一般に難易度が高い。相関や回帰の結果は交絡や共変動の影響を受けやすく，回帰係数や有意性だけで因果方向を断定することはできない。そこで本節では，観測データだけで因果を最終結論として“確定”するのではなく，前節で高精度に絞り込んだ重要因子群を対象に，一定のアルゴリズムと統一した判定基準に基づいて“上流（原因系）／下流（結果系）”の向きの候補を抽出し，SCN 除去に直接寄与し得る因子を検証優先度の高い順に整理することを目的として因果解析を導入する[9]。

本研究で扱う NGS 解析データは，微生物組成の偏り，離散性，非対称性などにより，誤差分布が正規分布から外れる傾向が強い。この点は，誤差の非ガウス性を仮定して因果方向を推定する LiNGAM（Linear Non-Gaussian Acyclic Model）[10] を適用する上で有利に働く。本節では，選抜済みの重要微生物群（および目的変数）に対して LiNGAM に基づく階層化（上流→下流の順序推定）を行い，さらに回帰に基づく依存関係の強さを可視化することで，SCN 除去に対する直接寄与の候補を整理する。

図８は，まず LiNGAM により重要微生物群および SCN 除去の因果順序を階層として推定し，その階層順に変数を並べ替えたうえで，階層化した各因子について“上流側の因子だけで条件付きに説明できるか（関係が残るか）”を調

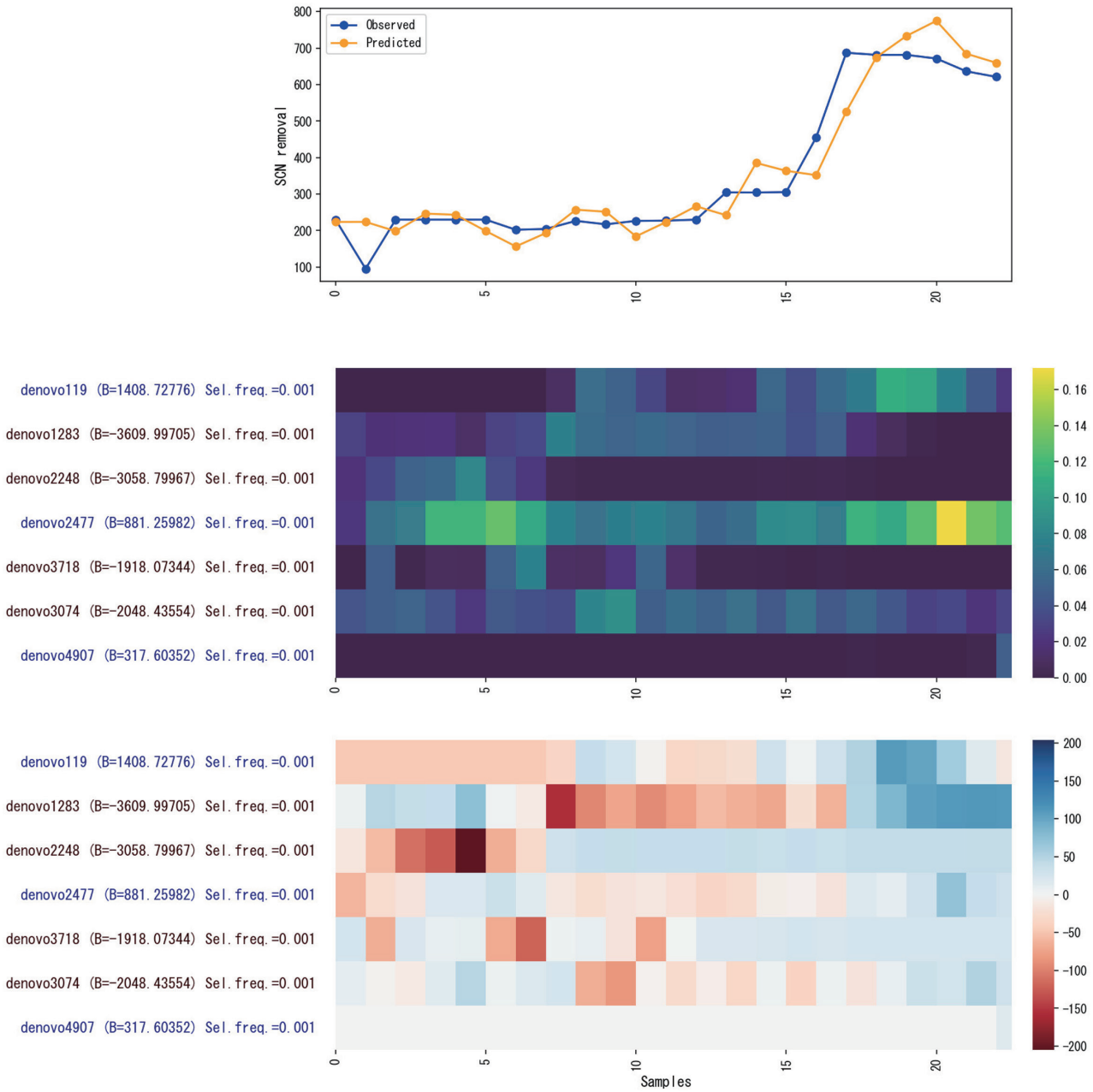

図 7　時系列予測と因子寄与の可視化

図 6 で得られた 7 因子モデルについて，上段に目的変数（SCN 除去）の実測値と予測値の時系列推移を示す。中段は各説明変数（重要微生物）の相対存在割合（サンプル内で正規化した値）のヒートマップ，下段はそれらの変動が予測値に与える寄与量（中心化した説明変数値×回帰係数）のヒートマップである。なお，寄与量の計算方法と色の解釈は図 4 の定義に従う。図中の Sel.freq. は Selection frequency（信頼度）の略。

Time-series prediction and visualization of factor contributions.

For the seven-factor model obtained in Fig. 6, the top panel shows the time-series trajectories of the observed and predicted values of the response variable (SCN removal). The middle panel is a heatmap of the relative abundances of each explanatory variable (important microbial taxa), normalized within each sample, and the bottom panel is a heatmap of their contributions to the predicted values (centered explanatory-variable values × regression coefficients). The contribution calculation and color interpretation follow the definitions in Fig. 4. “Sel. freq.” is an abbreviation for selection frequency (confidence).

べる目的で，線形重回帰を行い，標準誤差をロバスト化[11]して評価した結果を可視化したものである。ここで重要なのは，各行（目的変数）に対する回帰では，その因子より上流に位置する因子のみを説明変数として用いる点である。したがって図 8 では，斜線で示した対角（自己参照で回帰に入れない領域）を境に，左側（上流側）のセルのみが“説明変数候補（原因系候補）”として評価対象となり，右側（下流側）は構造上そもそも説明変数に入れないため，色が付かない（評価しない）。各セルは，列の因子を説明変数，行の因子を目的変数とする回帰における回帰係数の符号（正／負）と，ロバスト標準誤差に基づく p 値（有意性）を統合した Signed score[12] を表し，p 値が閾値未満の組合せのみを因果順序の有向辺候補として表示した。

図 8 の SCN 除去の行を横方向に見ると，上流側（対角よ

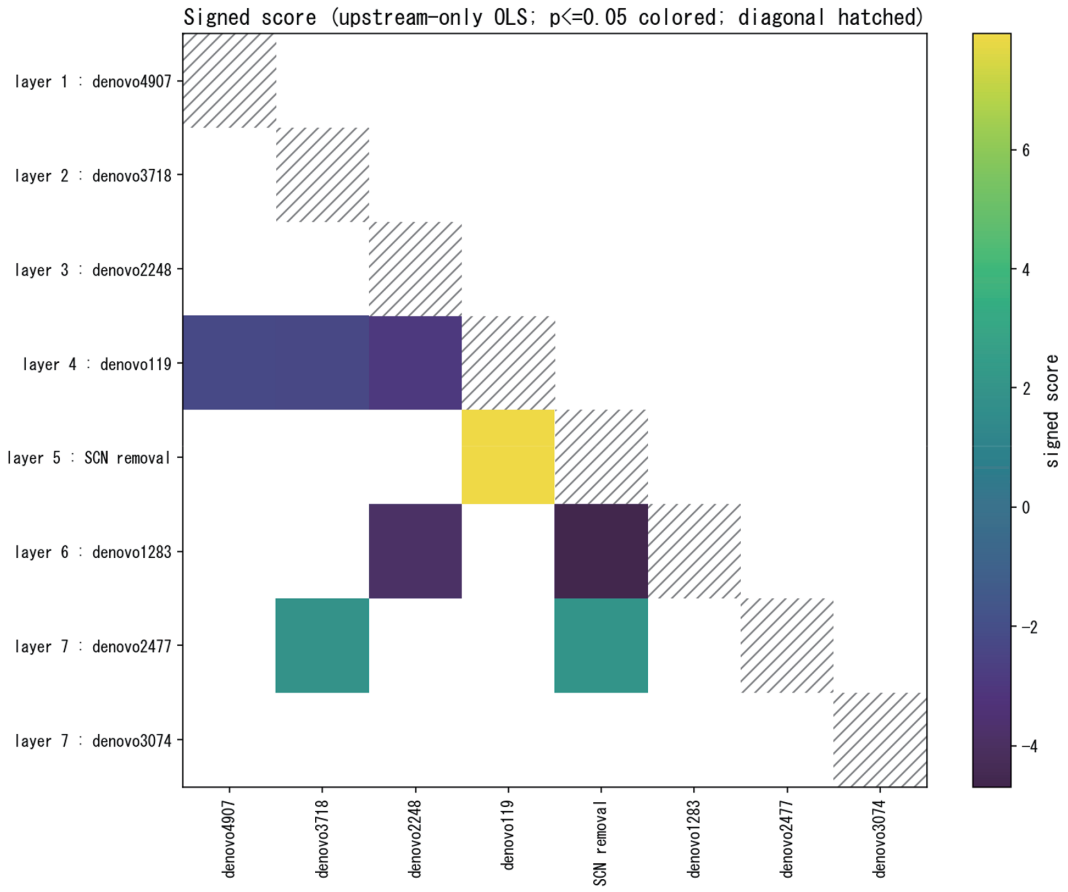

図 8　回帰係数の符号と有意度（ロバスト p 値）を統合したスコアに基づく依存関係の可視化
各行は“その行の変数を目的変数”とした線形重回帰の結果を表し，各列は説明変数を表す。ただし各行の回帰では，LiNGAM で推定した階層において当該変数より上流に位置する因子（対角より左側）のみを説明変数として採用したため，対角より右側（下流因子）のセルは評価対象外として白抜きで示す。色は回帰係数の符号（正／負）と統計的根拠の強さを同時に示すため，Signed score[9] を $S=\mathrm{sign}(\beta)\times(-\log10(p_{robust}))$ で定義した（β：回帰係数，p_{robust}：ロバスト p 値）。正の関係は正方向，負の関係は負方向の値として表現され，|S| が大きいほど統計的根拠が強いことを意味する。p_{robust} が閾値（0.05）未満の組合せのみを着色し，それ以外は白抜きとした。対角（自己回帰に相当）は斜線で示す。なお，回帰係数の比較可能性を確保するため，解析に用いた各変数は事前に標準化した。
Visualization of inter-variable dependencies based on a score integrating the sign of regression coefficients and statistical significance (robust p-values).
Each row shows the results of a multiple linear regression in which the variable in that row is treated as the response variable, and each column represents an explanatory variable. In each row-wise regression, however, only factors located upstream of the target variable in the hierarchy estimated by LiNGAM (i.e., cells to the left of the diagonal) were included as explanatory variables; therefore, cells to the right of the diagonal (downstream factors) are outside the scope of evaluation and are shown in white. To simultaneously represent the sign of the regression coefficient (positive/negative) and the strength of statistical evidence, we defined the signed score[9] as $S=\mathrm{sign}(\beta)\times(-\log10(p_{robust}))$ where β is the regression coefficient and p_{robust} is the robust p-value. Positive relationships are expressed as positive values and negative relationships as negative values, and larger |S| indicates stronger statistical evidence. Only pairs with $p_{robust}<0.05$ are colored; all others are left white. The diagonal (corresponding to self-regression) is indicated by hatching. To ensure comparability of regression coefficients, all variables used in the analysis were standardized in advance.

り左側）で統計的に有意な関係が残る因子として denovo119 が抽出される。これは，LiNGAM が与えた因果順序（上流→下流）に整合する形で，上流因子のみを同時に考慮してもなお denovo119 の関係が独立に残ることを意味し，SCN 除去に対する直接寄与因子（原因系）の優先候補として位置づけられる。加えて，denovo119 の行に着目すると，対角より左側に位置する複数の上流因子が同時に有意として現れており，上流側の情報が denovo119 に集約されていることが示唆される。すなわち denovo119 は，上流因子群の影響を受け取る“結節点”として振る舞い，そのうえで SCN 除去に対して独立な関係が残る点で，直接寄与因子候補としての重要性が高い。denovo119 は Chromatiales 目に属しており，同じ Chromatiales 目に分類される微生物を水処理プロセスから分離培養しチオシアン分解能を確認した報告（Oshiki et al.[13]）とも整合的である。

一方，denovo1283（Nitrosomonadales 目）は SCN 除去より下流（layer 6）に配置されており，図 8 では denovo1283 の行において SCN 除去が説明変数（上流因子）として現れ，負方向（負の Signed score）で有意に残る。これは，denovo 1283 が SCN 除去の原因側というより，SCN 除去の状態変

化に追随して応答する結果系として解釈することを支持する。さらに負方向で現れることは，SCN 分解系と硝化系（亜硝酸生成系）が競合・置換関係にあるという生態学的関係を反映している可能性が高い（Prosser et al.[14]）。

3. 結　　言

本報告では，コークス炉排水処理における重要微生物の同定に有効であった Lasso＋Bootstrap 法について，従来指摘されてきた課題，すなわち“信頼度（選択頻度）に対する閾値設定の指導原理が乏しく，解析者の主観が入り得る”点，および“ノイズの影響が大きいデータでは重要因子候補が多数残り，因子選択の安定性・再現性が低下し得る”点に対し，予測性能と偽陽性抑制の観点から一連の改善手法を提案した。

2.1 では，信頼度ランキングに基づいて説明変数を逐次追加し，予測性能（決定係数 R^2）が局所最大（あるいは最大）となる点で因子数を打ち切ることにより，重要因子群を定義する手順を提案した。これにより，信頼度の閾値を恣意的に定めるのではなく，“予測性能を最大化しつつ過学習を避ける”という評価軸に基づいて因子群を決定でき，信頼度利用の実務上の課題に対して一定の指導原理を与えた。

2.2 では，Model-X ノックオフの考え方を取り入れた END（Enhanced Noise Discriminative）-Lasso を導入し，統計的に生成したノイズ変数を“比較対象”として学習に同時投入することで，ノイズに起因する偽陽性を抑制しつつ重要因子候補の絞り込みを強化した。さらに逐次特徴選択と組み合わせることで，候補因子集合から予測精度を最大化する因子組合せを決定できる枠組みを示した。

2.3 では，前節までに高精度に絞り込んだ重要因子群を対象に，NGS 解析データが示す非ガウス性という特性を踏まえて LiNGAM を適用し，因果順序（上流→下流）を推定した。続いて，その順序に基づき，各因子について“上流因子のみを同時に考慮しても関係が残るか（条件付きで説明されるか）”をロバスト標準誤差に基づく多変量回帰（OLS）で評価し，因果順序と因子間関係を統合的に可視化した。これにより，観測データのみから因果を最終結論として断定するのではなく，統一した手順と判定基準の下で，上流（原因系）／下流（結果系）の向きの候補を整理し，SCN 除去に直接寄与し得る因子を検証優先度の高い順に提示できる枠組みを示した。

以上より，本報告で提案した手法群は，(i) 信頼度の絶対閾値設定に依存しない重要因子群の決定，(ii) ノイズの強いデータに対する偽陽性抑制と選択安定性の向上，(iii) 絞り込み後の重要因子群に対して“因果順序（上流→下流）”と“条件付きで残る関係（ロバスト回帰の有意性）”を組み合わせ，原因系候補の抽出と検証優先順位づけまでを一貫して支援する，という実務的要求に対応する解析体系を提供する。

なお，本手法群は水処理分野に限定されず，他の微生物群集データ（発酵，土壌，医療・衛生など）や，さらには微生物以外を含む膨大な候補因子から重要因子を抽出する必要がある分野（品質管理，プロセス運転，材料・化学プロセス，環境データ解析等）にも適用可能である。また，目的変数が水処理速度のような連続値に限らず，〇×のような 2 値データ（達成／未達，良／不良，合格／不合格など）を対象とする解析需要も想定されるため，評価指標・モデル（ロジスティック回帰等）を適切に選択することで同様の枠組みとして展開できる。今後は，2 値データを含む多様なデータ特性に対する実証，ならびに現場条件・運転指標との統合による介入可能な制御指針への展開を進めることで，より広範な社会課題の解決に資する統計解析手法としての確立が期待される。

参照文献

1) 福島寿和 ほか：日本製鉄技報．(417)，86 (2021)
2) 日本特許出願公告 7299485．2023 年 6 月 28 日
3) 川野秀一 ほか：スパース推定法による統計モデリング．初版．東京，共立出版，2018，168p
4) Efron, B. et al.: An introduction to the bootstrap. 1st ed. New York, Chapman and Hall/CRC, 1994, 456p
5) 日本特許出願公開 2025-171971．2025 年 11 月 20 日
6) Barber, R.F. et al.: The Annals of Statistics. 43, 2055 (2015)
7) Candès, E.J. et al.: Journal of the Royal Statistical Society: Series B (Statistical Methodology). 80 (3), 551 (2018)
8) 日本特許出願公開 2025-171281．2025 年 11 月 20 日
9) 特願 2024-181917．処理装置，処理方法，およびプログラム
10) Shimizu, S. et al.: Journal of Machine Learning Research. 7 (72), 2003 (2006)
11) MacKinnon, J.G. et al.: Journal of Econometrics. 29 (3), 305 (1985)
12) Severin, Y. et al.: Science Advances. 8 (44), eabn5631 (2022)
13) Oshiki, M. et al.: Microbes and Environments. 34 (4), 402 (2019)
14) Prosser, J. et al.: The prokaryotes. 4th ed. Heidelberg, Springer, 2014, p.901

福島寿和　Toshikazu FUKUSHIMA
先端技術研究所　環境基盤研究部
上席主幹研究員　博士(環境学)
千葉県富津市新富20-1　〒293-8511

中川淳一　Junichi NAKAGAWA
前　先端技術研究所　数理科学研究部
上席主幹研究員　博士(数理科学)

川野秀一　Shuichi KAWANO
九州大学　大学院数理学研究院　解析部門
教授　博士(機能数理学)

押木　守　Mamoru OSHIKI
北海道大学　大学院工学研究院　環境工学部門
准教授　博士(環境学)