統計数理(2004)
第 52 巻 第 2 号 275–279

特集「環境科学と統計科学の融合に向けて」
[総合報告]

環境科学への取り組みに対する期待*

北川 源四郎†

(受付 2004 年 3 月 17 日)

要　　旨

情報化社会の進展に伴って，大量データからの知識発見と予測が重要な課題となっている．一方，グローバル化は社会の不確実性を増大させ，リスクの適切な評価に基づく判断が重要になっている．これらの問題解決のためには，統計的モデリングが不可欠であり，ここに統計科学の現代的な役割がある．このように統計科学の役割がますます重要化する一方，複雑化し巨大化する現実の問題は，統計科学の方法の新たな展開の契機となった．本稿では，環境科学への取り組みにおいて，統計科学が貢献できることと，統計科学の発展の契機となりうることの両側面から考えてみる．

キーワード：大量データ，予測，知識発見，不確実性，リスクの管理，統計的モデリング．

1. 情報化社会と環境科学において求められるもの

地球環境の問題は今や人類共通の課題となった．いうまでもなく，この問題の解決のためには世界全体の取り組みが必要であり，国際協力と国際競争が避けられない．このような状況のなかで，我が国が国際的な発言力を強めていくためには，情報化社会に対応した説得力を持った科学的方法の確立が必要であり，そのためには，世界最先端のデータ解析技術が必要である．我が国の科学技術政策においても，これを視野に入れた組織的取り組みを行うべきである．

20 世紀後半に急速に発達した計測・通信・計算の技術は，情報化社会を創出した．これに伴って，地球科学，生命科学，ファイナンス，マーケティングなどの分野では，大量データが蓄積し，その有効利用が喫緊の課題となっている．環境科学においても状況は全く同じである．前世紀の一時期，統計学は厳密な設計の下で得られた少数データから，精密な推論を行うことを目指し多くの成果を挙げてきた．しかしながら，20 世紀の後半に出現した情報化社会は統計科学をめぐるこれまでの状況を一変させた．現在ではむしろ，氾濫する大量データの中から，有用な情報を抽出し，知識発見や予測へと至るための方法論の確立が求められている．そのためには，様々な目的で取得された多種類のデータを統合することにより，目的に即して有用な情報を抽出し，新しい知識を発見したり将来の予測を行うための合理的な方法の研究・開発が求められている．

20 世紀の科学技術は，基本的にはニュートン力学に代表される確定的な世界観に基づいていたといえるであろう．しかしながら，情報化された社会は，あらゆる分野でグローバル化をも

† 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7

* 本稿の内容は ISM シンポジウム（2002 年 8 月 19 日）における講演,「環境科学と統計的モデリング── 予測と発見の科学を目指して──」の一部に基づくものである．

たらした．その結果，環境問題に限らず，金融・経済，保険，生命，疫学，防災，安全性などの様々な領域で社会的問題が顕在化し，不確実性の増大を引き起こしている．これらの問題の解決のためには，不確実性を正面から捉え，リスクの存在を十分考慮した判断や行動決定が不可欠となる．ここにおいて，不確実性を伴う現象をデータに基づいてモデル化するための手段を与える統計科学が不可欠となる．

2. 統計科学の役割

科学的研究においては何らかの形で，データによって代表される現実との対比が要求される．データ取得の方法，データに基づく統計的モデリングの方法，そして得られたモデルに基づく推論・予測・知識発見およびリスク評価の方法を与える統計科学は，科学的研究の「言語」と社会的問題へ科学的に接近するための手段を与える．情報化社会における環境科学の研究に要求される課題はまさに統計科学が目指すものそのものといえる．

統計的モデルの重要な役割は，データに基づいて将来の変動を予測し，データから重要な情報を抽出することである．そのためには，対象の特性と目的に応じて適切にモデルを構成することが不可欠であり，それはデータの持つ情報と対象に対する知識，理論，経験，これまでのデータ，そして分析の目的，言い換えれば事前情報を適切に併合してモデルを構成することによって実現される．とくに，環境データの解析においては，様々な目的で観測されたデータや補助情報，知識や制約などの情報が混在するので，実際のモデリングにおいては，様々なレベルの知識とデータなどの異種情報の統合が不可欠となる．この異なる種類の情報統合は，ベイズモデル，特に時系列の場合には状態空間モデル，の利用によって実現できる（Akaike (1980)）．したがって，ベイズモデルや状態空間モデルの構成法および関連する計算法の実用化がますます重要な課題となってくる．

しかしながら，環境科学の問題解決には統計科学のさらなる発展が必要である．環境科学においては，地球規模で様々な要因が互いに関連しあい，時間変動する現象のモデリングが必要となる．これは環境データのモデリングでは，大量データの解析が必要となることを意味するが，それはデータ数と変量の増大による計算量の増加を意味するだけに止まらない．統計モデルは常に現実の複雑さと推論の限界のはざまで構成されているので，大量データの存在は，より複雑でより柔軟なモデルの導入を要求することになる．MCMC や逐次モンテカルロ法などのベイズモデルや状態空間モデルに関する統計計算法は，複雑なモデルの実用化のための重要な手段を与えたといえる．

このような統計的モデルは，前提とする事前情報や知識によって様々なものが得られるので，利用するモデルの良し悪しが情報処理の結果に直結する．したがって，モデル評価を客観的に評価するモデル評価基準が重要である．AIC（赤池情報量規準）がこのための客観的な評価基準を与えることは統計的モデリングの実用化において重要である（小西・北川（2004））．

3. 時系列解析における当面の具体的な課題

環境科学，特に地球環境予測への適用を想定するとき，大量データの処理とモデリングに向けて取り組むべき問題が見出される．近年，樋口らの研究グループ（新世紀重点研究プロジェクト「先端的四次元大気海洋陸域結合データ同化システムの開発と高精度気候変動予測に必要な初期値化・再解析データセットの構築」）はデータ同化の観点から，関連する様々な問題に取り組んでいる．本節ではそのような具体的課題の例として，筆者が関係する時系列解析の分野において問題となる，超高次元時系列モデルの実用化，超多変量時系列データからの情報抽出・知識発見と複雑・巨大なシステムの予測・シミュレーションの方法の確立などのやや技術的な

問題について考えてみる事にする．

3.1 超高次元時系列モデルの実用化

環境科学のデータ解析では，多数の時空間データを関連づけて解析することが必要である．したがって，これらの変数を統合して得られる統計モデルは超高次元となるのでパラメータ推定や状態推定のための様々な工夫が必要となる．

超高次元 AR モデルのパラメータ推定法および変数選択法． 定常時系列解析において多変量 AR モデルは，時系列の予測と制御，フィードバックシステムの解析，パワー寄与率の計算などに利用されその有効性が実証されている．しかしながら，環境データの解析においては，しばしば数百変量から 1000 変量以上の高次元データのモデリングが必要になる．超高次元データのモデリングにおいては，無用な変数を排除するために変数選択が重要であるが，k 変量 m 次の多変量 AR モデルは k^2m 個以上の未知数を持ち，AR 係数の推定自体も困難となるので，予測誤差の相関行列を順次推定する新しい推定法の利用が考えられる．

超高次元カルマンフィルタのアルゴリズム． 時系列のモデリングにおいて，状態空間モデルは様々な種類の情報を統合する強力な手段となる．これを利用して，非定常時系列のモデリング，時変パラメータをもつモデルの推定などに利用されてきた(赤池・北川(1994，1995))．状態空間モデリングにおいては重要な状態推定の問題は，線形・ガウス型の状態空間モデルに対してはカルマンフィルタによる効率的な推定が可能である(片山(2000))．しかし，超多変量の時系列に対してはカルマンフィルタの計算さえも困難となる．したがって，超高次元状態空間モデルに対する効率的・実用的なフィルタリングのアルゴリズムの開発が必要である．

高次元一般状態空間モデルのフィルタリング． 環境問題に関連するモデリングにおいては，非線形性，非ガウス性，異常値などが混在し，通常の状態空間モデルでは適切に取り扱えないことが多い．このような場合には，一般状態空間モデルの利用が有効であることが知られている(Kitagawa (1987，1996))．ただし，この一般状態空間モデルに対してはカルマンフィルタではよい推定値が得られないので，非ガウス型フィルタなどの数値的方法の利用が必要である．近年はモンテカルロフィルタなどの粒子近似に基づく計算法や自己組織型モデルによるパラメータ推定の方法が開発されているが，これを高次元状態空間モデルに適用できるようにすることも必要である．

3.2 超多変量時系列データからの情報抽出・知識発見の方法の確立

多変量時系列の解析においては多変量 AR モデルに基づくパワー寄与率が，フィードバックが存在する複雑な多変量システムの因果分析において重要な役割を果たす．超多変量 AR モデルの推定法が実用化されると，直ちにパワー寄与率を適用することが考えられる．しかしながら，従来のパワー寄与率はイノベーション項の独立性が仮定されており，環境問題においては現実的な仮定とはいえない．したがって，ノイズ成分間に相関がある場合にも適用可能な新しい解析法の開発が望まれる．これに関しては Tanokura and Kitagawa (2003)は拡張されたパワー寄与率の定義とその計算法を提案し，より一般のシステムの解析法の可能性を開拓している．今後は，定常成分に限らず，非定常成分の因果分析の方法開発が望まれる．

3.3 複雑・巨大なシステムの予測・シミュレーションの方法の確立

気象予報，地震予測，経済予測などにおいて，値による予測から分布による予測へという転換が定着しつつある．環境予測や状態推定においても，確率分布の計算が不可欠である．複雑な高次元非線形システムの予測分布の計算は，一般には極めて困難であるが，シミュレーションの繰り返し計算により比較的簡単にその近似値を求めることができる．

特に統計数理研究所に導入されたブートストラップシステムは物理乱数に基づくシミュレー

ションを並列計算機上で実現するので，並列シミュレーションや予測分布の近似を極めて効率的に実現できる．このシステムを有効に利用する計算法の確立も重要な課題と思われる．

参 考 文 献

Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

赤池弘次，北川源四郎(編)(1994)．『時系列解析の実際 I』，朝倉書店，東京．

赤池弘次，北川源四郎(編)(1995)．『時系列解析の実際 II』，朝倉書店，東京．

片山　徹(2000)．『応用カルマンフィルタ(新版)』，朝倉書店，東京．

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association*, **82**, 1032–1063.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space model, *Journal of Computational and Graphical Statistics*, **5**, 1–25.

小西貞則，北川源四郎(2004)．『情報量規準』，朝倉書店，東京．

Tanokura, Y. and Kitagawa, G. (2003). Generalized power contribution to detect correlated noise sources, Research Memorandum, No. 902, The Institute of Statistical Mathematics, Tokyo.

Proceedings of the Institute of Statistical Mathematics Vol. 52, No. 2, 275–279 (2004) 279

Statistical Approach to Environmental Science

Genshiro Kitagawa

The Institute of Statistical Mathematics

With the development of various information technologies in measurement, communication and computation, prediction and knowledge discovery based on massive data sets have become crutial problems. However, "globalization" has increased the uncertainty of society , and decisions based on proper evaluation of risk has become important. To solve these two important problems, statistical modeling is indispensable, and it also reveals some challenges in research on statistical science.

Key words: State space models, information criterion.